Blueprint Labs

Discussion Paper #2022.17

Methods for Measuring School Effectiveness

Joshua Angrist
Peter Hull
Christopher R. Walters

December 2022

MIT Department of Economics
77 Massachusetts Avenue, Bldg. E53-390
Cambridge, MA 02139

National Bureau of Economic Research
1050 Massachusetts Avenue, 3rd Floor
Cambridge, MA 02138

NBER WORKING PAPER SERIES

METHODS FOR MEASURING SCHOOL EFFECTIVENESS

Joshua Angrist
Peter Hull
Christopher R. Walters

Working Paper 30803
http://www.nber.org/papers/w30803

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022

We thank Sarah Cohodes, Parag Pathak, and research staff at the Massachusetts Department of Elementary and Secondary Education for helpful comments. Jerray Chang, Nidhaan Jain, and Hellary Zhang provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w30803

Methods for Measuring School Effectiveness
Joshua Angrist, Peter Hull, and Christopher R. Walters
NBER Working Paper No. 30803
December 2022
JEL No. C11,C26,I20,I21,I24

ABSTRACT

Many personal and policy decisions turn on perceptions of school effectiveness, defined here as the causal effect of attendance at a particular school or set of schools on student test scores and other outcomes. Widely-disseminated school ratings frameworks compare average student achievement across schools, but uncontrolled differences in means may owe more to selection bias than to causal effects. Such selection problems have motivated a wave of econometric innovation that uses elements of random and quasi-experimental variation to measure school effectiveness. This chapter reviews these empirical strategies, highlighting solved problems and open questions. Empirical examples are used throughout.

Joshua Angrist
Department of Economics, E52-436
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
angrist@mit.edu

Peter Hull
Department of Economics
Box B, Brown University
Providence RI 02912
and NBER
peter_hull@brown.edu

Christopher R. Walters
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

1 Introduction

A wide range of personal and policy decisions turn on perceptions of school quality. Families choose schools and neighborhoods by balancing perceived school effectiveness against other factors, like housing costs and commute times. Questions of school quality also drive high-stakes policy decisions around school closure, restructuring, and expansion. In response to the demand for such information, achievement-based measures of public school effectiveness have proliferated. Such measures include "school report cards" distributed by some states and districts, as well as school quality ratings published by private organizations like GreatSchools.org.

How can school quality be reliably measured? This chapter reviews recent econometric strategies for estimating school effectiveness, defined as the causal effect of attending a particular school or set of similar schools (like charter schools) on student outcomes. Efforts to gauge school quality must confront the fundamental challenge of selection bias: school-to-school comparisons can reflect student ability and family background as much as or more than school effectiveness. Economists have devised an array of solutions to such bias; increasingly, these use elements of randomness in modern school assignment schemes to devise convincing natural experiments.

Our review begins with the relatively straightforward matter of how to gauge the effectiveness of a single school or school sector that offers seats by lottery. For instance, Angrist et al. (2010) study effects of the first Knowledge is Power Program (KIPP) charter school in New England. In principle, lotteries for seats at a single oversubscribed charter school identify the causal effects of attendance at this school. This identification strategy is implemented using the (perhaps conditionally) randomized offer of seats at the school in question as an instrumental variable (IV) for school attendance.

Although conceptually straightforward, complications arise in implementing school-lottery IV, even to estimate the quality of a single school or sector. Our discussion of lottery basics covers problems related to covariates, waiting lists, and the delays between random assignment, school enrollment, and measurement of outcomes. We also discuss the regression-discontinuity (RD) analog of single-school lotteries where students are admitted according to whether a test score clears a cutoff, instead of by conditionally random assignment. The leading example here is a selective-enrollment "exam" school like those studied in Abdulkadiroğlu et al. (2014). Finally, the single-school IV setting is used to review methods for the estimation of school effects on test score distributions, as in Angrist et al. (2016a).

Of course, many questions of school quality involve more than one sector or school. An analyst aspiring to measure multiple causal effects must establish a common counterfactual to ensure the estimates are true apples-to-apples comparisons. The identification of multi-school and multi-sector models is facilitated by centralized school assignment, as in the Boston, Denver, New Orleans, and New York City school districts (to name a few). Centralized assignment schemes randomly assign a good share of seats at most schools in the centralized district. Random assignment in such districts is conditional: different students are assigned seats at a given school at different rates determined by preferences and priorities. Abdulkadiroğlu et al. (2017) and Abdulkadiroğlu et al. (2021) derive the assignment probabilities arising in a centralized match, and show how the vector of assignment rates lays the foundation for causal inference with multiple sectors or schools. We synthesize and illustrate practical lessons from this work.

Interest in school effectiveness clearly predates the advent of centralized assignment, however. Conventional value-added models (VAMs) gauge school effectiveness using regression to control for lagged outcomes and other covariates. Because VAM estimates for individual schools tend to be noisy, economists have long deployed empirical Bayes (EB) methods to reduce sampling variance.[1] More recently, Angrist et al. (2016b, 2017, 2021) develop a suite of EB methods using centralized assignment to measure school effectiveness and to reduce selection bias in conventional VAM estimates. These methods aim to balance the relative precision of conventional VAMs with the bias reduction and model validation afforded by centralized assignment. New VAM models and methods are illustrated here with applications to schools in Massachusetts, Denver, and New York City. As in Angrist et al. (2021), these applications show how centralized assignment can shed light on school effectiveness even for schools where no seats are randomly assigned.

We structure this chapter as follows. Section 2 reviews the basic IV framework as it applies to quality measures for a single school or sector. Section 3 extends the framework to cover models with heterogeneous effects, measures of effectiveness over multiple sectors and years, and identification strategies using RD-style seat assignment. Section 4 describes the use of centralized assignment to jointly estimate the quality of multiple distinct sectors. A leading example here is an analysis of different types of charter schools. Section 5 discusses regression-controlled VAM estimation, outlines an EB framework for the analysis of individual school quality, and shows how quasi-experimental admissions or assignment variation can be used to validate and improve on conventional VAMs. We conclude by highlighting remaining challenges and research frontiers in the measurement of school effectiveness.

2 School Lottery Basics

2.1 Single-School Effects

A large and expanding empirical literature uses randomized admission lotteries to gauge the effects of K-12 schools on academic outcomes in the United States. Examples include studies of charter schools (Hoxby and Murarka, 2009; Angrist et al., 2010, 2012, 2013, 2016a; Abdulkadiroğlu et al., 2011; Dobbie and Fryer, 2011, 2013, 2015; Clark et al., 2015; Davis and Heller, 2019; Cohodes et al., 2021; Setren, 2021), school vouchers (Chingos and Peterson, 2015; Mills and Wolf, 2017; Abdulkadiroğlu et al., 2018), small schools (Bloom and Unterman, 2014), magnet schools (Engberg et al., 2014), boarding schools (Curto and Fryer, 2014), and aspects of school choice (Cullen et al., 2006; Hastings et al., 2009; Deming, 2011, 2014; Deming et al., 2014).[2] We begin by reviewing econometric methods for the simplest question this work considers, that of the effectiveness of a single school.

[1] Raudenbush and Bryk (1986) are the first to apply EB methods in this context.

[2] Outside the US, lottery-based school evaluations include Angrist et al. (2002), Lee et al. (2014), Zhang (2014), Behaghel et al. (2017), Lee and Nakazawa (2021), Oosterbeek and de Wolf (2021), and Romero and Singh (2021).

Suppose a researcher is interested in the effects of attendance at a charter school in the KIPP network. KIPP has been often in the public eye: network schools are emblematic of the *No Excuses* approach to public education, a widely replicated urban charter model that features a long school day, an extended school year, selective teacher hiring, extensive data-driven feedback for teachers, student behavior norms, and a focus on traditional reading and math skills. Not long ago, there was only one such school in New England—the KIPP middle school in Lynn, Mass. How good is it? Angrist et al. (2010, 2012) gauge KIPP Lynn effectiveness using data on KIPP lottery applicants.

The lottery strategy uses IV to identify the causal effect of a Bernoulli treatment, $D_i \in \{0, 1\}$, here indicating KIPP enrollment for student i. Let $Y_i(1)$ denote this student's potential 6th grade achievement level if she attends KIPP Lynn, and let $Y_i(0)$ denote her achievement otherwise. The observed outcome for this student, Y_i, is one or the other of these two potential outcomes depending on D_i. We can thus write:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i. \end{aligned}$$

To focus on the problem of selection bias, we initially postulate a constant causal effect, $Y_i(1) - Y(0) = \beta$, for all i. Under constant effects, observed outcomes can be described by writing:

$$Y_i = \mu + \beta D_i + \varepsilon_i, \tag{1}$$

where $\mu \equiv E[Y_i(0)]$ and $\varepsilon_i = Y_i(0) - \mu$. The term ε_i can be thought of as a composite measure of student ability, family background, and motivation for schoolwork. We call this measure "ability" for short.

The core challenge in estimating β comes from the fact that ability and KIPP attendance are likely correlated. KIPP students may be especially motivated or come from more educated families than the typical urban student (Rothstein, 2004), such that $E[\varepsilon_i \mid D_i = 1]$ exceeds $E[\varepsilon_i \mid D_i = 0]$. In this case, the causal parameter β in equation (1) is unlikely to coincide with the slope coefficient in a regression of Y_i on D_i or, equivalently, the difference in conditional means of Y_i with D_i switched on and off. Selection bias likely causes comparisons of mean achievement by KIPP enrollment status to exceed β:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \beta + \underbrace{E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0]}_{\text{Selection bias}} > \beta.$$

Multivariate regression estimates that control for family background and pre-application (i.e., lagged) achievement ameliorate, but do not necessarily eliminate, this sort of selection bias, as unobservable ability differences may remain given such controls.

IV using randomized admission lotteries answers the challenge of selection bias. Suppose n applicants apply for $m < n$ 6th grade seats at KIPP. Let dummy variable Z_i indicate applicants offered a seat. We assume that lotteries are both fair and consequential. Formally, this means,

Assumption 1A. *Lottery offers are independent of student ability:* $\varepsilon_i \perp\!\!\!\perp Z_i$.

Assumption 1B. *Lottery winners are more likely to enroll than losers:* $E[D_i|Z_i = 1] > E[D_i|Z_i = 0]$.

Assumption 1A is plausible because lottery offers are randomly assigned and so independent of student ability, though 1A also requires offers be unrelated to outcomes through channels other than charter attendance. The latter is an exclusion restriction. Assumption 1B is plausible because lottery offers open the door for enrollment: while not all lottery winners enroll, and while some lottery losers may find their way in through other channels, enrollment is strongly predicted by offers.

This pair of assumptions makes Z_i a valid instrument for D_i in equation (1). Since the instrument here is Bernoulli, the IV estimand is a Wald (1940)-type ratio of differences in means:

$$\beta_{IV} \equiv \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]} = \beta. \tag{2}$$

Identification of β follows from equation (1) and the facts that Assumption 1A implies $E[\varepsilon|Z_i] = 0$, while Assumption 1B ensures the denominator of (2) is nonzero.

The numerator and denominator of (2) can be computed as the slope coefficients from bivariate regressions of achievement and charter attendance on Z_i:

$$Y_i = \gamma + \rho Z_i + u_i, \tag{3}$$

$$D_i = \psi + \pi Z_i + v_i. \tag{4}$$

Conventionally, the first of these equations is called the *reduced form* while the second is the corresponding *first stage*. We sometimes use these terms to refer to the slope coefficients, ρ and π, rather than the equations containing them. Equation (2) shows that $\beta_{IV} = \rho/\pi$, which coincides with the causal parameter β under Assumptions 1A and 1B.

2.2 Control for Assignment Risk and Covariates

Admissions lotteries are typically run every year. Most are run separately for different entry grades and for siblings of students already enrolled, among other special groups. Such multi-lottery scenarios can be seen as a naturally-occurring analog of the stratified randomized controlled trials (RCTs) used to gauge drug efficacy. In a stratified RCT, where subjects are treated at different rates in different strata, treatment assignment is independent of potential outcomes only within strata.

The signal feature distinguishing lottery strata is the conditional probability of an offer, a feature we call *assignment risk*. For instance, assignment risk among applicants might be 0.9 when a school first opens and demand is weak such that virtually all applicants get in. But assignment risk might decline to 0.5 or less in later years as the school becomes established and popular. Likewise, siblings are typically offered seats at a much higher rate than non-siblings in a given year. A group of students with the same assignment risk is said to constitute a *risk set*. Differences in assignment risk are a possible source of selection bias even in a stratified RCT. We therefore control for assignment risk in all but the simplest, single-stratum, lottery-based research designs.

Within risk sets, the single-lottery IV framework applies. Let $R_i \in \{1, ..., K\}$ encode the identity of applicant i's risk set, with dummies $R_{ik} = 1[R_i = k]$ indicating i is in the k^{th} set. The conditional Wald estimand for applicants in risk set k is:

$$\beta_{IV,k} \equiv \frac{E[Y_i|Z_i = 1, R_i = k] - E[Y_i|Z_i = 0, R_i = k]}{E[D_i|Z_i = 1, R_i = k] - E[D_i|Z_i = 0, R_i = k]}. \tag{5}$$

If Assumptions 1A and 1B hold conditional on $R_i = k$, then $\beta_{IV,k}$ captures the causal effect of charter attendance for applicants in this group. In practice, risk-set-specific estimates based on the sample analog of $\beta_{IV,k}$ are likely to be noisy. But we can aggregate these conditional estimates into a single, more precise, summary estimate.

A two-stage least squares (2SLS) estimator with a full set of risk set controls conveniently aggregates the set of $\beta_{IV,k}$ in a single weighted average. Augmenting the causal model described by (1) and the first stage equation described by (4) with risk control leads to a 2SLS setup described by:

$$Y_i = \beta D_i + \sum_{k=1}^{K} \delta_k R_{ik} + \eta_i, \tag{6}$$

$$D_i = \pi Z_i + \sum_{k=1}^{K} \tau_k R_{ik} + \upsilon_i. \tag{7}$$

Parameters δ_k in the risk-controlled causal model, (6), can be viewed as coefficients from a regression of $\mu + \varepsilon_i$ in (1) on the set of R_{ik}, with associated residual η_i. First stage equation (7) likewise controls for risk by including the R_{ik} as regressors (as a rule, a 2SLS first stage includes the controls appearing in the causal model with which it's associated).

2SLS uses the first-stage fitted values generated by (7) to instrument D_i in (6).[3] Kolesár (2013) shows that this 2SLS estimand, characterized by a dummy endogenous variable with saturated control for discrete

[3] The reduced form equation for this system can be written:

$$Y_i = \rho Z_i + \sum_{k=1}^{K} \gamma_k R_{ik} + \xi_i,$$

where the γ_k are reduced-form risk-set effects. Because this model is just-identified (i.e., the number of instruments equals the number of variables to be instrumented), 2SLS coincides with an indirect least squares estimator that divides OLS estimates of reduced-form coefficient ρ by estimates of the corresponding first-stage coefficient, π, in (7).

covariates, captures a weighted average of within-risk-set IV coefficients that can be written:

$$\beta_{2SLS} = \sum_{k=1}^{K} \left[\frac{\omega_k \pi_k p_k (1-p_k)}{\sum_l \omega_l \pi_l p_l (1-p_l)} \right] \beta_{IV,k}. \tag{8}$$

Here, $\omega_k = Pr[R_{ik} = 1]$ is the share of applicants in risk set k; $\pi_k = E[D_i|Z_i = 1, R_{ik} = 1] - E[D_i|Z_i = 0, R_{ik} = 1]$ is the corresponding first stage for these applicants; and $p_k = Pr[Z_i = 1|R_{ik} = 1]$, so $p_k(1-p_k)$ is the conditional offer variance. The weights on each $\beta_{IV,k}$ are non-negative if Assumption 1B holds for each lottery so that $\pi_k > 0$ for all k. This weighting scheme is likely to generate precise estimates of β by giving more weight to strata with more students, stronger first stages, and more instrument variation.[4]

An alternative 2SLS estimator adds interactions between the offer instrument and risk set dummies to the list of excluded instruments, leading to an over-identified model with a fully saturated first stage.[5] 2SLS with a saturated first stage generates a different weighted average that replaces π_k with π_k^2 in (8), as detailed in Angrist and Imbens (1995). Under constant effects and homoskedasticity of the residuals in (6), 2SLS estimates of this over-identified model are efficient in the sense of yielding the (asymptotically) most precise estimate that can be computed given the information at hand. On the other hand, 2SLS estimates of heavily-overidentified models are likely to be subject to more finite-sample bias than estimates of just-identified models (Andrews et al., 2019; Angrist and Kolesár, 2021).

Covariate Control

Beyond risk set controls, 2SLS estimators for school quality often incorporate additional covariates describing students' background. Although not required to eliminate selection bias, non-risk covariates typically increase the precision of 2SLS estimates. A 2SLS setup with covariates is described by:

$$Y_i = \beta D_i + \sum_{k=1}^{K} \delta_k R_{ik} + X_i'\mu + \eta_i, \tag{9}$$

$$D_i = \pi Z_i + \sum_{k=1}^{K} \tau_k R_{ik} + X_i'\psi + \upsilon_i. \tag{10}$$

Covariate vector X_i might include applicant demographic characteristics like race, sex, and free-lunch status, as well as baseline test scores measured prior to the lottery in which the applicant participates.

Assuming lottery offers are randomly assigned within risk sets, Z_i and X_i are uncorrelated conditional on assignment risk. This implies the coefficient on X_i should be zero in a regression of Z_i on risk set dummies and X_i. Controlling for X_i therefore leaves the 2SLS estimand unchanged. This fact is a version of the Frisch-Waugh-Lovell theorem for 2SLS. To the extent that controlling for X_i reduces the residual variance of outcomes, however, 2SLS estimates of models with X_i included can be expected to be more precise than estimates of models with X_i omitted.

[4] Goldsmith-Pinkham et al. (2022) study the efficiency of such regression-based weighting schemes.

[5] This first stage is saturated because it includes a parameter for each value of $E[D_i \mid Z_i, R_i]$.

Multiple Schools in a Single Sector

Importantly, the risk set idea extends to analyses of sectors covering more than one school. Suppose the city of Lynn has two KIPP middle schools, KIPP A and KIPP B. Applicants for 5th grade seats may apply to one or both of these charter schools. In a multi-school/single-sector analysis, applicants are coded as having attended KIPP if they enroll in either school; no attempt is made to identify causal effects of KIPP A and B separately. In a single-sector analysis, we pursue a common KIPP treatment effect. Moreover, in this example, the KIPP effect in Lynn is also a charter-school sector effect since the city of Lynn has no non-KIPP charters.

Multi-school scenarios open the door to alternative IV strategies. We might, for example, use two offer dummies as instruments with one for each school. With only one endogenous variable indicating any KIPP attendance, 2SLS estimates are over-identified by two instruments. Each instrument in this case generates distinct assignment risk. Often, however, a just-identified single-instrument estimator is appealing—both for pedagogical reasons, and because models with many instruments (in more elaborate real-world applications) can be subject to finite-sample bias.

The typical just-identified setup uses a single any-offer instrument. With two schools, the any-offer instrument indicates applicants receiving offers from one or both schools; assignment risk is the probability of the union of the two underlying offer events. If, for instance, the A and B lotteries are independent, and if applicant i has risk $p_i(A)$ at school A and $p_i(B)$ at school B, then the risk of a KIPP offer is $p_i(A) + p_i(B) - p_i(A)p_i(B)$. Applicants who apply to only one or the other school have either $p_i(A) = 0$ or $p_i(B) = 0$. When assignment risk takes on three values determined solely by the identities of schools applied to, there are therefore three risk sets. We next section illustrates this scenario in an analysis of Massachusetts urban charter schools.

2.3 Massachusetts Urban Charter Effects

An investigation of Massachusetts urban charter middle school effects highlights key elements of a basic lottery analysis of school effectiveness. The sample used for this purpose is that analyzed in Angrist et al. (2013). The 2SLS setup here includes both risk set controls and baseline (non-risk) covariates. Risk set controls consist of indicators for all combinations of charter schools applied to, separately by application year. The treatment effect of interest is the effect of attending one of Massachusetts urban charter middle schools, mostly in Boston. This investigation compares schools in the urban charter sector to all other public schools, both traditional and charter.

Table 1 reports summary statistics for this sample. Column (2) displays baseline (4th grade) characteristics for 6,038 students applying for fifth- or sixth-grade entry between 2002 and 2011 at one of 9 urban charters with available lottery records. The lottery sample keeps only the earliest recorded application year for each student and excludes applicants receiving guaranteed seats (e.g., applicants with enrolled siblings), resulting in a set of students subject to random lottery admission. For comparison, characteristics of the full Massachusetts urban school district population over the same period are displayed in column (1).

The summary statistics show some notable differences between lottery applicants and the broader student population. In particular, lottery applicants are more likely to be Black, are less likely to be classified as English language learners, and have somewhat higher baseline test scores. While both groups are below the state average in fourth grade, charter applicants have math and English scores roughly 0.1 standard deviations (σ) higher than the urban average (here all test scores are normalized to mean zero and standard deviation one in each grade and year for the state of Massachusetts). These differences illustrate the potential for selection into charter attendance, highlighting the value of lottery-based experiments.

Balance and Attrition

Two empirical checks probe the validity of the assumptions underlying a lottery-based analysis of school effectiveness. The first is a covariate balance check: with random lottery offers, baseline characteristics of lottery winners and losers should be similar. Column (3) of Table 1 displays coefficients from regressions of fourth-grade characteristics on a lottery offer dummy, defined as an indicator equal to one if a student receives an offer from any charter school, controlling for assignment risk. Students are coded as offered if they received either an offer on lottery day or a later offer while on a waiting list. Distinctions between these offer types are discussed in Section 3.4, below.

Balance coefficient estimates, reported in column (3) of Table 1, show differences between winners and losers that are uniformly small and statistically insignificant, consistent with random assignment of charter offers within risk sets. A joint test of the null hypothesis that all characteristics are balanced fails to reject at conventional levels ($p = 0.69$), similarly building empirical support for Assumption 1A within risk sets.

Even where offers are randomly assigned, a lottery analysis can be compromised by non-random differences in the likelihood of sample attrition for winners and losers. For example, some students may exit the public school system when not offered a charter seat, but attend a charter when offered the opportunity to do so. Such selective attrition can change the composition of lottery winners and losers remaining in the public system, potentially generating selection bias. It's therefore worth examining differences in followup rates between lottery winners and losers. Selection bias is unlikely to be a problem when followup rates for these two groups are similar (Lee, 2009).

The bottom of Table 1 reports the coefficient from a regression of an indicator for outcome test score availability in the first post-lottery year on an offer indicator, in a model with risk set controls. The overall followup rate in the Angrist et al. (2013) sample is 80.1 percent, with little difference in followup between winners and losers. This suggests attrition is not a major concern in this case. Engberg et al. (2014) and Abdulkadiroğlu et al. (2018) discuss bounds on school lottery treatment effects in cases where differential attrition is worrying.

2SLS Estimates

A well-documented 2SLS analysis reports first-stage estimates as well as 2SLS estimates of the causal effect of interest. The ingredients of a basic charter lottery analysis are laid out in Table 2. The treatment variable here is an indicator for charter attendance in the school year beginning in the fall of the year an application was submitted; the outcome is from tests taken at year's end (in 5th or 6th grade). Results in the table are from models that include risk-set dummies, student demographic characteristics, and baseline (fourth-grade) math and English scores.

Boston charter offers boost charter enrollment rates by nearly 60 percentage points, a large first stage estimate that can be seen in the first pair of columns in Table 2. Although unsurprising, this result establishes the validity of first-stage Assumption 1B within risk sets. 2SLS estimates, reported in the first row of column (2) in the table, show charter attendance boosts math scores by an impressive 0.45σ. Non-charter students in this sample score 0.32σ below the state mean. One year in a Boston charter is therefore predicted to boost math achievement to a level above the state average. A corresponding OLS estimate, displayed in column (1), produces a charter coefficient of 0.33σ. Evidently, despite concerns of positive selection bias, the uninstrumented model understates the IV effect. One possible explanation for this is treatment effect heterogeneity, an issue we turn to next.

3 Lottery IV: Implementation Details and Extensions

3.1 Heterogeneous Effects

The causal model motivating Table 2 specifies a constant causal effect of charter attendance, at least within lottery risk sets. In practice, of course, charter effects may differ for different applicants. Individually-varying charter effects are defined by potential outcomes: $\beta_i = Y_i(1) - Y_i(0)$. Different students may likewise respond differently to the offer of a charter seat. A heterogeneous first stage can be described by a set of potential treatments, $D_i(1)$ and $D_i(0)$, indicating i's charter enrollment status when $Z_i = 1$ and $Z_i = 0$, respectively.

In a world of heterogeneous potential outcomes, IV captures causal effects on charter-lottery compliers. This is formalized by the local average treatment effects (LATE) interpretation of the Wald estimand. Specifically, Imbens and Angrist (1994) and Angrist et al. (1996) show:

$$\beta_{IV} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)].$$

This result is derived assuming independence, exclusion, the existence of a first stage, and monotonicity, defined as follows:

Assumption 2A. *Independence/exclusion:* $(Y_i(1), Y_i(0), D_i(1), D_i(0)) \perp\!\!\!\perp Z_i$.

Assumption 2B. *First stage:* $E[D_i|Z_i = 1] > E[D_i|Z_i = 0]$.

Assumption 2C. *Monotonicity:* $D_i(1) \geq D_i(0)\ \forall i$.

Assumption 2A adapts our earlier Assumption 1A to the heterogeneous treatment effects framework, while Assumption 2B aligns with our previous Assumption 1B.[6] The novel requirement is Assumption 2C, which states that the lottery offer instrument must weakly increase charter enrollment for all students (not just on average). This monotonicity restriction seems sensible, since it is hard to imagine a scenario in which lottery wins make attendance less likely.

The LATE Theorem partitions the population of charter applicants. "Never-takers" are students who decline to attend charter schools even when offered, so that $D_i(1) = D_i(0) = 0$. This might be a student whose lottery participation is due to the enthusiasm of a parent rather than their own. "Always-takers" are students who attend charter even without an offer, so that $D_i(1) = D_i(0) = 1$. Students in this category might apply repeatedly or otherwise find a way in. Charter lottery "compliers" are those who attend if and only if they receive offers, implying that $D_i(1) > D_i(0)$ (i.e. $D_i(1) = 1$ and $D_i(0) = 0$). The LATE Theorem implies that β_{IV} delivers the average causal effect of charter attendance for compliers. With multiple lotteries, equation (8) implies that 2SLS with risk set controls identifies a weighted average of lottery-specific LATEs. Estimates like those in Table 3 should therefore be interpreted as measuring causal effects for applicants induced to attend charter schools by randomized offers. These effects may be higher or lower than the overall average effect of charters, depending on how compliers compare to others.

3.2 Characterizing Compliers

Complier Covariate Means

Who are charter lottery compliers? Remarkably, while individual compliers are not coded as such in any data set (since we never see $D_i(1)$ and $D_i(0)$ for the same student i), complier characteristics can be described. Tools for this are detailed in Abadie (2002, 2003).

Complier analyses build on the fact that, by virtue of Assumption 2C, the population contributing to an IV analysis contains only always-takers, never-takers, and compliers. Moreover, unlike compliers, some always and never-takers are identifiable: students with $D_i = 0$ and $Z_i = 1$ must be never-takers, while those with $D_i = 1$ and $Z_i = 0$ must be always-takers. The other two combinations of D_i and Z_i involve mixtures of compliers and these other groups: the population with $D_i = Z_i = 1$ is a mixture of always-takers and compliers, while the group with $D_i = Z_i = 0$ is a mixture of never-takers and compliers. The relative sizes of the three groups are identified, since the never-taker share is given by the fraction of offers that are declined, the always-taker share is given by the fraction of non-offered students that attend charters, and the complier share equals the size of the first stage. We can therefore back out characteristics of lottery compliers from the mixture distributions combined with the observed information about always and never-takers.

Abadie (2002) implements this logic with a simple 2SLS procedure for characterizing compliers, described

[6] Assumption 2A combines independence and exclusion, distinct assumptions in the LATE framework.

by:

$$g(X_i, Y_i) \times 1\{D_i = d\} = \psi_d + \gamma_d 1\{D_i = d\} + v_{id}, \tag{11}$$

$$1\{D_i = d\} = \phi_d + \pi_d Z_i + e_{id}, \quad d \in \{0, 1\}. \tag{12}$$

Here $g(X_i, Y_i)$ is any function of student baseline characteristics X_i and post-lottery outcomes Y_i. Setting $d = 1$ in (11) and (12) means that we are using Z_i as an instrument for D_i in an IV procedure with $g(X_i, Y_i)$ multiplied by D_i as the outcome. Setting $d = 0$ means we use Z_i to instrument $(1 - D_i)$ in an equation with $g(X_i, Y_i)(1 - D_i)$ as the outcome. Under Assumptions 2A, 2B, and 2C (and the maintained assumption that Z_i is independent of X_i), Abadie (2002) shows that these unconventional IV procedures recover characteristics of treated and untreated compliers:

$$\gamma_d = E[g(X_i, Y_i(d)) | D_i(1) > D_i(0)], \qquad d \in \{0, 1\}.$$

This result has a number of useful implications and applications. First, by setting $g(X_i, Y_i) = X_i$, the IV procedure produces the average of any predetermined covariate X_i for lottery compliers, facilitating comparisons of compliers with other groups on observable dimensions. Note that since X_i is unaffected by the charter treatment, both IV coefficients γ_1 and γ_0 recover the mean X_i for compliers in this case, so these two parameters should be equal. Complier characteristics can be summarized by reporting an estimate of either parameter, or an average of the two estimates. It is straightforward to show that the difference between these two estimates is proportional to the difference in X_i by offer status used as a balance check in Table 1. We can therefore think of such balance checks as equivalently checking covariate balance across treatment and control groups in the hidden complier RCT embedded in a lottery with partial compliance.

Table 3 illustrates this method for the Angrist et al. (2013) Massachusetts urban charter applicant sample. Columns (1) and (2) show the two available estimates of complier means in the untreated and treated states for a list of baseline characteristics. These results come from 2SLS estimation of (11) and (12) augmented with additive risk set controls; per equation (8), we can interpret the estimates as weighted averages of risk set-specific complier means. As expected given the balance checks in Table 1, the treated and untreated complier estimates are very similar for all characteristics. Column (3) shows a more efficient single estimate of each complier mean, constructed by stacking the data for the two systems in (11) and (12) and conducting 2SLS estimation imposing a common coefficient γ across equations, clustering standard errors by student.[7]

Columns (4) and (5) of Table 3 present mean characteristics for always and never-takers for the purposes of comparison with the compliers. As noted above, characteristics of these groups can be computed based on the $(D_i = 1, Z_i = 0)$ and $(D_i = 0, Z_i = 1)$ cells of data, respectively. For comparability with the 2SLS weighting used to compute complier means, we estimate always-taker means by regressing $X_i D_i(1 - Z_i)$ on

[7] In principle an even more efficient estimate of the complier mean could be formed by conducting three-stage least squares (3SLS) estimation of the stacked system, exploiting the covariance structure across equations (Zellner and Theil, 1962).

$D_i(1-Z_i)$ with additive risk set controls, and we estimate never-taker means by regressing $X_i(1-D_i)Z_i$ on $(1-D_i)Z_i$ with risk set controls. These procedures automate aggregation across risk sets with regression-based weights, as with the IV estimands in columns (1)-(3).

The comparisons in Table 3 reveal differences between behavioral response types for Massachusetts urban charter lotteries. Compliers are less likely to be Black and more likely to be white or Hispanic than always- or never-takers, and are more likely to be classified as English language learners. Always-takers are the lowest-achieving group as measured by baseline test scores. As discussed further in Section 3.7, complier characteristics provide a partial guide to the external validity of a set of lottery-based IV estimates.

Complier Potential Outcome Distributions

A second application of the Abadie (2002) method produces marginal potential outcome distributions for compliers.[8] By setting $g(X_i, Y_i) = Y_i$ and $d = 1$ in equations (11) and (12), we obtain an IV coefficient γ_1 equal to the complier average of the charter outcome $Y_i(1)$. Similarly, setting $g(X_i, Y_i) = Y_i$ and $d = 0$ gives a γ_0 equal to the complier mean of the non-charter outcome $Y_i(0)$. This demonstrates that the levels of mean potential outcomes are identified for compliers, not just the difference between them (i.e. the LATE). These mean potential outcomes can be compared to the mean $Y_i(0)$ for never-takers and mean $Y_i(1)$ for always-takers, which are directly observed. Such comparisons serve as the basis for tests for selection into treatment (Angrist, 2004), and as inputs into parametric modeling approaches that extrapolate from LATE to predict other treatment effect parameters (Brinch et al., 2017; Kline and Walters, 2019).

In fact, the full marginal distributions of both potential outcomes are identified for compliers—not just the means. By setting $g(X_i, Y_i) = 1\{Y_i \leq y\}$ for a constant y and each value of d, we obtain the complier cumulative distribution functions of $Y_i(1)$ and $Y_i(0)$ evaluated at y.[9] This in turn implies that quantile treatment effects (QTEs) are identified for compliers. Complier QTEs can be computed either by inverting the complier CDFs or via weighted quantile regression, an approach detailed in Abadie et al. (2002).

Densities are often easier to interpret than CDFs, especially in a graphical analysis. Complier densities at a point y can be estimated by setting $g(X_i, Y_i) = \frac{1}{h}K\left(\frac{Y_i - y}{h}\right)$ in equation (11), where $K(\cdot)$ is a symmetric kernel function maximized at zero and h is a bandwidth that shrinks to zero asymptotically. Estimating equations (11) and (12) with this choice of $g(\cdot)$ produces estimates of complier potential outcome densities evaluated at y. In an application to private school voucher lotteries, Abdulkadiroğlu et al. (2018) propose a bandwidth selection procedure that adapts Silverman (1986)'s rule-of-thumb approach to yield a bandwidth appropriate for complier density estimation.[10]

[8] The joint distribution of complier treatment effects $Y_i(1) - Y_i(0)$ is generally not point identified. Frandsen and Lefgren (2021) show how bounds on this joint distribution can nevertheless be formed using charter lotteries.

[9] Complier potential outcome CDFs obtained this way need not be weakly increasing. Decreasing CDFs signal a violation of the underlying independence, exclusion, or monotonicity assumptions. Huber and Mellace (2015) and Kitagawa (2015) use this idea to test instrument validity.

[10] Silverman's rule-of-thumb for a Gaussian kernel $K(\cdot)$ sets $h = 1.06 \times N^{-1/5}\sigma$, where N is the sample size and σ is the standard deviation of the outcome. Abdulkadiroğlu et al. (2018) plug in consistent estimates of these quantities for compliers, estimating the number of compliers as the first stage times the total sample size and the first two moments of complier potential outcomes by setting $g(X_i, Y_i) = Y_i$ and $g(X_i, Y_i) = Y_i^2$, in equation (11).

Figure 1 illustrates this complier density estimation technique for the Angrist et al. (2013) Massachusetts urban charter applicant sample. A notable result in Angrist et al. (2013) is that urban charters generate much larger test score gains for non-white students than for white students, reducing racial achievement gaps. Figure 1 reports complier math score densities separately for white and Black students, with the density of $Y_i(0)$ on the left and $Y_i(1)$ on the right. These results come from 2SLS estimation of equations (11) and (12) with risk set controls added to each equation. By equation (8), the estimates capture weighted averages of within-risk-set complier densities. The top panel shows densities of baseline scores from the year prior to the lottery. As expected, score distributions are similar for treated and control compliers at baseline. The densities of baseline scores for Black students are shifted left relative to those of white students, indicating a large racial achievement gap. A bootstrap Kolmogorov-Smirnov (KS) test rejects racial equality of baseline score distributions for both untreated and treated compliers ($p < 0.01$).[11]

The subsequent panels of Figure 1 demonstrate that urban charter attendance eliminates the racial achievement gap for lottery compliers. The treated densities on the right reveal a marked convergence of the Black distribution toward the white distribution after the lottery, so that these two distributions are virtually indistinguishable by seventh grade. The KS test fails to reject equality of $Y_i(1)$ distributions for Black and white compliers ($p = 0.94$) in this grade. In contrast, the distributions on the left show a persistent achievement gap for compliers randomized into traditional public schools. The racial gap in $Y_i(0)$ distributions in seventh grade is similar to the gap at baseline, and racial equality of the non-treated outcome distributions is rejected ($p < 0.01$).

3.3 Multiple Years

The analysis above is limited to outcomes in a single post-lottery grade, with a treatment variable that measures school attendance in the year following the lottery. When outcomes are available for more than one grade, they can be combined in a pooled analysis. A 2SLS setup that stacks grades can be described by:

$$Y_{ig} = \beta D_{ig} + \sum_{k=1}^{K} \delta_k R_{ik} + X_{ig}'\mu + \eta_{ig}, \tag{13}$$

$$D_{ig} = \pi Z_i + \sum_{k=1}^{K} \tau_k R_{ik} + X_{ig}'\psi + \upsilon_{ig}, \tag{14}$$

where Y_{ig} is student i's outcome in grade g and covariate vector X_{ig} includes grade and calendar year effects along with other baseline characteristics. Since assignment risk is fixed for a given student regardless of grade observed, risk controls needn't vary by grade.

Multi-grade models typically introduce a grade-varying treatment to reflect exposure to schools or sectors of interest. The Abdulkadiroğlu et al. (2011) and Angrist et al. (2013) charter-school studies implement this. Specifically, endogenous variable D_{ig} is defined as years enrolled in charter between the lottery and grade

[11] See the figure notes for details on this testing procedure.

in which an outcome is observed. This adjusts for differences in exposure due to reapplication or dropout, under the assumption that we need only know total time enrolled in a charter to satisfy the relevant exclusion restriction. We assume away, for instance, differences in timing details such as which particular grades were attended at a charter. Assuming total duration of school attendance mediates the effects of the offer, the 2SLS estimand in equation (13) captures an Average Causal Response (ACR) of the sort defined in Angrist and Imbens (1995). ACR generalizes LATE to the case of treatments with variable intensity. In this case the ACR measures a weighted average of per-year impacts of school attendance for students whose enrollment choices are shifted by the lottery offer.

Estimates generated by a stacked multi-grade setup for Massachusetts urban charter applicants appear in columns (3)-(6) of Table 2. These estimates were computed by adding post-lottery outcomes through eighth grade to the first-year sample. The estimate in column (4) use the same dummy endogenous variable as the single-year analysis. The 2SLS estimate of 0.58σ exceeds that in column (2), but the magnitude of this estimate is complicated by the fact that applicants experience up to four years of charter attendance. The estimates in column (6) were computed using years of exposure as the endogenous variable. The first stage for this specification is slightly greater than one, while the resulting ACR estimate is 0.31σ, implying—on average across grades and lotteries—each year of charter enrollment boosts math scores by roughly one-third of a standard deviation for compliers. As in the single-grade analysis, OLS estimates in columns (3) and (5) are positive but smaller than corresponding 2SLS estimates, suggesting a modest downward bias in regression-adjusted comparisons of charter and non-charter students.

3.4 Coding Lottery Offer Instruments

We have so far ignored matters of timing. In practice, lottery-admitting schools often make a first round of initial offers, with students not initially offered a seat placed on a waiting list with students ordered by lottery number. As some initial offers are declined, offers are made down the waiting list. Information on initial and waitlist offers can be used to construct multiple instruments for charter enrollment.

A natural two-instrument strategy combines an initial offer dummy with a dummy indicating waitlist offers using 2SLS. Since take-up for initial and waitlist offers may differ (if, for example, waitlisted students enroll elsewhere before receiving an offer), over-identified 2SLS may generate more precise estimates than a just-identified model using a single offer dummy. The Abdulkadiroğlu et al. (2011) analysis of Boston charter schools shows modest efficiency gains from a two-instrument setup. It's worth noting, however, that the LATE interpretation of over-identified 2SLS estimates requires a stronger monotonicity assumption than that typically invoked with single-instrument IV (Mogstad et al., 2021).

de Chaisemartin and Behaghel (2020) discuss use of waitlist instruments in settings with few students per lottery. When schools target class size, instruments constructed from waitlist offers in lotteries with few applicants can be correlated with potential outcomes. This problem arises from the fact that with a class size target, the last student who receives an offer must be a complier, resulting in over-representation of compliers among offered students. de Chaisemartin and Behaghel (2020) propose a weighted IV estimator that ameliorates the bias in IV estimates using waitlist offers. Initial offer instruments also circumvent this problem by using a pre-determined offer cutoff that does not depend on takeup.

In principle, data on randomly-ordered lottery lists can be used to construct more precise estimates while avoiding the use of realized waitlist offers. Let L_i denote the order assigned to applicant i in a charter lottery, where applicants are randomly ordered from $\{1, ..., \bar{L}\}$. An initial offer instrument is an indicator for L_i below a fixed cutoff C. More generally, the efficient function of L_i to use as instrument is the expected charter enrollment rate at each lottery number, which can be computed given a model for offer takeup. Suppose, for instance, that a school plans to make offers until it enrolls a class size of C students, that offer take-up is independent across students, and that there are no always-takers. Then the total number of offers equals C plus a negative binomial random variable with C successes and success probability equal to the offer take-up rate, π. This implies the likelihood that a student at lottery list position $L_i = \ell$ ultimately enrolls in this charter school is:

$$Pr[D_i = 1|L_i = \ell] = \pi \times \left[1 - 1\{\ell > C\}\left(1 - \frac{\int_0^\pi u^{\ell-C-1}(1-u)^{C-1}du \times (\ell-1)!}{(\ell-C-1)!(C-1)!}\right)\right].$$

This formula reveals that a student with $L_i \leq C$ is assured of getting an offer, in which case the probability of attendance is the compliance rate π; for students with $L_i > C$, the second factor captures the probability that at least one seat remains to be offered at their position on the wait-list. We have yet to see this optimal IV approach applied to lottery quasi-experiments. In practice, problems with the use of waitlist offers are likely to matter little when lottery sizes are large.

3.5 Multi-Sector Models

The lottery framework extends to models and methods that capture multiple sector effects in one go. We might be interested, for example, in distinguishing the effects of KIPP charter schools from those of schools belonging to other networks. After introducing the multi-sector framework, Section 4 outlines methods that exploit centralized, algorithmic assignment for identification and estimation of sector effects. Section 5 extends this with an overview of value-added models that allow distinct causal effects for each of many schools, without regard to sector. While similar in broad strokes, each area of analysis raises unique conceptual and implementation challenges.

Counterfactual Destinies for Lottery Compliers

The analysis of urban charters sketched in Section 2.3 raises an important conceptual question: "compared to what?" Among applicants to schools in a particular charter sector, lottery losers might attend traditional public schools, charters belonging to another sector, or one of a number of exam schools, to name a few alternative sectors. It's helpful, therefore, to characterize the distribution of enrollment across sectors to be compared. As in Abdulkadiroğlu et al. (2014) and Chabrier et al. (2016), we refer to this as the distribution of *counterfactual destinies.*[12]

Data from the Cohodes et al. (2021) study of new and veteran charter schools in Boston illustrate the destinies idea. This analysis estimates and compares effects of newly-opened charters (in the wake of a 2010 ballot initiative lifting the Boston charter cap) with effects of older schools. In Massachusetts charter school vernacular, only "proven providers" are permitted to open a new school; new schools associated with these are called "expansion campuses" in the study. The analysis here also considers effects of other charters, unaffected by expansion, and pilot schools, another autonomous Boston school model.[13]

Table 4 summarizes counterfactual destinies for compliers among applicants to each of these charter school types in the post-reform period. These estimates are computed via 2SLS estimation of equations (11) and (12) with $d = 0$ for applicants to a particular charter type, coding Z_i and D_i based on offers and enrollment at that charter type, and setting $g(\cdot)$ equal to an indicator for an enrollment in a specific sector (controlling for risk sets, as always). Column (1) shows that while 53% of compliers who do not receive a lottery offer at parent campuses attend traditional public schools, another 27% enroll at expansion campuses. Similarly, among compliers in lotteries for other charters (neither parents nor expansions), 23% of lottery losers end up enrolled at an expansion charter. In contrast, column (3) shows few untreated compliers in expansion charter lotteries enroll in other charter types, so the counterfactual for this group is composed primarily of BPS district schools (either traditional public schools or pilots). These results thus establish the importance of understanding the composition of the counterfactual for interpreting lottery results at each charter type.

Multi-Sector 2SLS

The counterfactual attendance patterns documented in Table 4 motivate an analysis of multiple school sectors in a unified framework, rather than contrasting single sectors against a composite counterfactual as in Section 2.3. We initially approach this with a constant-effects causal model that describes the consequences of attending one of several different school types.

Suppose each student in a district attends a school in one of S sectors numbered from 0 to S, with sector 0 representing traditional public schools. We define a mutually exclusive and exhaustive set of dummy variables to represent enrollment in each sector, with $D_{is} \in \{0, 1\}$ indicating attendance in sector s and $\sum_{s=0}^{S} D_{is} = 1$ for each student. A causal model with sector-specific effects is then given by:

$$Y_i = \mu + \sum_{s=1}^{S} \beta_s D_{is} + \varepsilon_i.$$

[12] See Kline and Walters (2016) and Feller et al. (2016) for analyses of counterfactual destinies in the context of early-childhood programs.

[13] More specifically, the 2010 initiative allowed existing charter schools which met the definition of a proven provider to apply to increase their maximum enrollment. Schools which were granted increases may or may not have opened additional campuses, depending on current facility capacity.

The parameter β_s measures the effect of attending a school in sector s relative to the omitted sector of traditional public schools. As before ε_i represents unobserved student heterogeneity that may be related to sector enrollment choices.

Now suppose we have a set of lotteries for enrollment in each sector. As in Section 2, assume there are K mutually exclusive lottery groups and let $R_{ik} = 1$ indicate that student i participates in lottery k. These lottery groups should be viewed as corresponding to all combinations of school-specific lotteries that a student might enter. Let Z_{is} denote an indicator equal to one if student i receives at least one offer from a school in sector s. Note that when students can apply to multiple lotteries, the Z_{is} are not necessarily mutually exclusive, as students may receive offers from multiple sectors. Extending equations (9) and (10) to the multi-sector setting results in the following system of equations with multiple endogenous variables:

$$Y_i = \sum_{s=1}^{S} \beta_s D_{is} + \sum_{k=1}^{K} \delta_k R_{ik} + X_i'\mu + \eta_i, \tag{15}$$

$$D_{is} = \sum_{m=1}^{S} \pi_{ms} Z_{im} + \sum_{k=1}^{K} \tau_{ks} R_{ik} + X_i'\psi_s + \upsilon_{is}, \quad s \in \{1, ..., S\}. \tag{16}$$

Paralleling Section 2, we can view the δ_k and μ in (15) as coefficients from a projection of $\mu + \varepsilon_i$ on risk set indicators and covariates, and the S first stage equations defined in (16) are projections of the sector attendance indicators on lottery offers along with risk sets and covariates. Two-stage least squares estimation proceeds by fitting each first stage equation by OLS, then running (15) by OLS after substituting in first-stage predicted values $\hat{D}_{is}$. By a simple extension of the above, this 2SLS procedure will recover consistent estimates of the sector effects β_s provided the offers Z_{is} are independent of ability η_i within risk sets and induce sufficient attendance variation for each sector. It is straightforward to extend this setup to stack outcomes across multiple grades as in Section 3.3.

Table 5 reproduces Cohodes et al. (2021)'s estimates of the effects of several Boston charter school types on math scores. This analysis treats parent campuses, expansion campuses, and other non-expansion charters as separate sectors, and also distinguishes between attendance before and after the expansion reform. Following the approach described in Section 3.3, the analysis stacks scores from all observed post-lottery grades for lottery applicants, and codes the endogenous variables D_{is} as the number of years spent in each sector. The instruments are indicators for initial and waitlist offers to each sector type, as described in Section 3.4, and the model controls for risk set indicators for the intersection of all school-by-year specific admission lotteries.

2SLS estimates of this multi-sector model show substantial treatment effects of attendance at both parent and expansion campuses on the order of one-third of a standard deviation per year. The effects of parent campuses are similar before and after the expansion reform, suggesting that parent schools' effectiveness was not diluted by expansion to new locations. Effects of other non-expansion charters are positive but smaller than those of the parents selected for expansion, indicating that the state of Massachusetts designated more effective schools for expansion. The inclusion of each of these charter types in a single 2SLS model means we can interpret the impact of each as relative to the same traditional Boston public schools benchmark.

Two additional features of this multi-sector approach are of note. First, multi-sector models like (15) generally rely on the assumption of constant effects of each sector across students. The LATE result of Imbens and Angrist (1994) does not apply to models with multiple endogenous variables, and in general a causal interpretation of 2SLS estimates from such models requires strong restrictions on either effect heterogeneity or behavioral responses to the instruments (Behaghel et al., 2013; Kirkeboen et al., 2016; Bhuller and Sigstad, 2022). Second, and similarly, the multi-sector model described here treats all schools within the same sector as equally effective. Heterogeneity in school quality within sectors complicates the interpretation of the estimates and creates the potential for exclusion restriction violations. For example, a student who switches from one expansion charter to another in response to a change in lottery offers experiences no change in the endogenous variables in equation (15), but may experience a change in scores if the two schools are of differing quality, thereby violating exclusion. Note that this sort of exclusion violation is a potential issue even for single-sector evaluations that code treatment as attendance at any charter in the sector, including charters without lotteries.

These issues motivate an approach that extends the sector model further to allow each individual school to have its own causal "value-added." However, the lottery methods described so far cannot be straightforwardly applied to such a model if only a subset of schools hold oversubscribed lotteries, because lotteries do not generate enough instruments to identify the effects of all schools. We return to the topics of individual school value-added and lottery undersubscription in Section 5.

3.6 Admission Discontinuities as Local Lotteries

A closely-related research design to the lottery methods described above leverages discontinuities in admissions rules based on admission tests or other criteria. For example, highly-selective exam schools in Boston, New York, and elsewhere require an admission test and admit students with a high enough score. We can view randomized lotteries as a special case of this sort of admission rule with a randomly-assigned admission score—as discussed in Section 3.4, lottery offers are distributed to students whose positions on a randomly-ordered list fall below a threshold. When the admission score is not randomly assigned, students above and below the admission threshold will not generally be comparable. This problem is solved with regression discontinuity (RD) methods that zero-in on a small neighborhood of the admission threshold, isolating comparisons of similar students who end up just above or below the threshold by chance.[14]

We introduce the admission RD approach by returning to the potential outcomes model of Section 3.1, this time describing the effects of exam school attendance. Let $Y_i(1)$ and $Y_i(0)$ denote student i's outcomes if she attends an exam school or a traditional public school, and let $D_i(1)$ and $D_i(0)$ represent i's attendance

[14] For other examples of evaluations of education programs derived from admission cutoffs, see Hoekstra (2009), Zimmerman (2014), Card and Giuliano (2016), Kirkeboen et al. (2016), Dustan et al. (2017), Heinesen (2018), Hastings et al. (2019), Zimmerman (2019), Anelli (2020), Sekhri (2020), Jia and Li (2021), Bleemer and Mehta (2022), Beuermann and Jackson (2022), and de Roux and Riehl (2022).

choices with and without an exam school admission offer. Instead of being randomly assigned as in the charter lottery context, the exam offer Z_i is assigned based on a cutoff c in an observed test score T_i:

$$Z_i = 1\{T_i \geq c\}.$$

We suppose that potential outcomes satisfy the following assumptions:

Assumption 3A. *Mean potential outcomes are smooth across the threshold:* $\lim_{t \to c-} E[Y_i(d)|T_i = t] = \lim_{t \to c+} E[Y_i(d)|T_i = t]$ *and* $\lim_{t \to c-} E[D_i(z)|T_i = t] = \lim_{t \to c+} E[D_i(z)|T_i = t]$ *for* $(d, z) \in \{0, 1\}^2$.

Assumption 3B. *Crossing the threshold increases enrollment:* $\lim_{t \to c+} E[D_i|T_i = t] > \lim_{t \to c-} E[D_i|T_i = t]$.

Assumption 3C. *Local monotonicity:* $Pr[D_i(1) \geq D_i(0)|T_i = c] = 1$.

These three assumptions are local versions of Assumptions 2A, 2B and 2C, applying only to students with scores in the immediate neighborhood of the admission threshold. The smoothness condition in Assumption 3A requires that students cannot precisely manipulate their scores in relation to the threshold, so that those just above and just below are similar in expectation. Under these conditions we can think of the admission threshold as defining a local randomized lottery for students with T_i close to c, and apply the lottery-based methods introduced earlier to this subpopulation. This idea is in keeping with recent work analyzing RD designs as local randomized trials (Cattaneo et al., 2016).

Under Assumptions 3A, 3B, 3C, a local version of the Wald ratio identifies the LATE for compliers at the admission threshold. Specifically, we have

$$\beta_{RD} \equiv \frac{\lim_{t \to c+} E[Y_i|T_i = t] - \lim_{t \to c-} E[Y_i|T_i = t]}{\lim_{t \to c+} E[D_i|T_i = t] - \lim_{t \to c-} E[D_i|T_i = t]} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0), T_i = c].$$

This expression can be seen as using the threshold indicator Z_i as an instrument for exam school enrollment D_i in the neighborhood of the threshold c.

Empirical implementation of the admission discontinuity design adapts the basic lottery 2SLS approach of Section 2 to the local lottery experiment.[15] Here, risk control involves the admission test score, T_i, known in RD vernacular as a *running variable*. Applicants with running variable values far from the cutoff have degenerate risk: they are either treated or not with probability one. Applicants with running variable values close to the cutoff may or may not clear it.

The simplest risk control strategy in this setting is parametric: polynomial functions of the running variable included as covariates adjust for the relationship between running variables and outcomes in the absence of treatment. In practice, however, a nonparametric strategy that looks only at applicants near the relevant cutoffs is often more convincing. Formal motivation for the local or nonparametric approach comes

[15] RD-2SLS strategies, where the instrument is an indicator for cutoff clearance, are sometimes called "fuzzy" RD designs. This contrasts with "sharp" RD designs where the treatment of interest is deterministic at the cutoff.

from the fact that, in a limiting sense made precise in Abdulkadiroğlu et al. (2021), the limiting probability of being offered a seat equals one-half in a shrinking bandwidth or interval around the cutoff.

Parametric and nonparametric RD estimation strategies can be described by the following 2SLS setup:

$$Y_i = \mu + \beta D_i + (1 - Z_i) f(T_i - c; \delta_0) + Z_i f(T_i - c; \delta_1) + \eta_i, \tag{17}$$

$$D_i = \psi + \pi Z_i + (1 - Z_i) f(T_i - c; \tau_0) + Z_i f(T_i - c; \tau_1) + \upsilon_i. \tag{18}$$

These equations replace the risk set indicators in (6) and (7) with a smooth function of the running variable $f(t; \delta)$ with parameter δ, satisfying $f(0; \delta) = 0$. In practice this function is typically a polynomial, so that $f(t; \delta) = \sum_{k=1}^{K} \delta_k t^k$. The parameters determining the polynomial coefficients are allowed to differ on each side of the threshold and in the first and second stage equations.

Parametric RD strategies use all or most of the sample of interest to compute 2SLS estimates of this model, usually with flexible running variable controls. Nonparametric RD downweights or removes observations farther from the cutoff and uses a less flexible running variable control (typically linear). The latter requires a choice of bandwidth to determine the sample size and weights to use in nonparametric estimation. The problem of how best to choose the bandwidth has stimulated substantial and ongoing theoretical work (see, e.g., Imbens and Kalyanaraman (2011) and Calonico et al. (2014)).

Figure 2, taken from Abdulkadiroğlu et al. (2014), plots RD first-stage and reduced form relations for New York City (NYC)'s highly selective exam schools. Panel A shows NYC exam school enrollment rates for applicants to three exam schools, Brooklyn Tech, Bronx Science, and Stuyvesant, as a function of the distance of a student's admission score to each school's admission cutoff. The figure reveals large jumps in enrollment across the cutoff, indicating a strong first stage for exam school attendance. Panel B shows corresponding reduced form impacts on Regents math standardized test scores. Regents scores are smooth through exam school admission cutoffs, revealing that exam school attendance has no impact on test scores for compliers. These zero impacts occur in spite of enormous differences in the level of achievement between exam and traditional public schools. This illustrates the power of discontinuity-based experiments to distinguish causal effects from selection bias.

3.7 External Validity

The results of this section demonstrate that lottery-based research designs identify average treatment effects for compliers, a well-defined and interpretable subpopulation. Lottery-based estimates are also relevant for evaluating certain policy reforms. Kline and Walters (2016) show that absent externalities and spillovers the LATE is the policy-relevant parameter for a marginal increase in the number of available seats among lottery applicants. It is often of interest to ask whether the external validity of the lottery LATE extends to other subpopulations or policy changes. At least four forms of external validity are worth considering.[16]

[16] See List (2021) and List et al. (2021) for related discussion of external validity and scaling of education programs.

First, among lottery applicants, effects for lottery compliers may differ from effects for always-takers and never-takers. While treatment effects for these other groups of applicants are not identified, the methods of Section 3.2 can be used to assess whether their observed characteristics or levels of potential outcomes differ from those of compliers. This kind of analysis can provide a sense of whether compliers are representative of the full population of applicants. It's worth noting that in many school lotteries always-takers may be rare or absent, since it may be difficult for students to enroll in a school without receiving an offer. With such one-sided non-compliance LATE is equal to the effect of treatment on the treated (TOT) among applicants, a traditional target parameter in the program evaluation literature (Bloom, 1984).

Second, effects for lottery applicants may differ from effects for students who choose not to apply to the lottery. Table 1 showed that characteristics of applicants and non-applicants differ in the Massachusetts urban charter example, suggesting that treatment effects may differ as well. To extrapolate to the population of non-applicants, it is useful to leverage other sorts of experiments that shift the composition of the lottery applicant pool. In one such application to Boston charter schools, Walters (2018) combines instruments based on distance to charter schools with randomized lotteries in a generalized Roy model framework (Roy, 1951; Eisenhauer et al., 2015). Intuitively, since students in the immediate neighborhood of a charter school are more likely to apply, those who apply from close by are less selected than those who apply from far away. Under the assumption that distance is as good as randomly assigned conditional on other observed characteristics, variation in the lottery LATE by distance can therefore be used to tease out the relationship between the propensity to apply and treatment effects. The results of Walters (2018) suggest that charter applicants are negatively selected on achievement gains, so that treatment effects may be even larger if charter schooling is expanded to new populations. Consistent with this finding, Cohodes et al. (2021) show that Boston charters continued to produce large achievement gains after a reform that resulted in an applicant pool more representative of Boston as a whole. Along similar lines, Abdulkadiroğlu et al. (2016) combine lottery records with an alternative research design based on charter takeovers of traditional public schools to show similar effects for lottery applicants and students "grandfathered" into charter schools.

Third, schools where lottery records are available may differ from non-lottery schools. For example, popular schools with more applicants than seats may be more effective than less popular schools, or schools with the administrative capacity to retain organized lottery records may be more effective.[17] Non-experimental estimates reported in Abdulkadiroğlu et al. (2011) and Angrist et al. (2013) suggest that oversubscribed charter schools in Massachusetts are more effective than schools without lotteries, and Baude et al. (2020) show that more effective charter schools gain market share over time in Texas. On the other hand, Abdulkadiroğlu et al. (2018) report that private schools with declining enrollment are more likely to opt into a lotteried voucher program, and Abdulkadiroğlu et al. (2020) show that school popularity is weakly related to school effectiveness among New York City high schools. Abdulkadiroğlu et al. (2014) and Dobbie and Fryer (2014) show limited effects of highly-sought-after exam schools in Boston and New York. It's therefore

[17] This idea is a version of the "site selection bias" studied by Allcott (2015).

not clear that we should expect oversubscribed schools to be more effective in general.

Finally, the presence of lotteried school choice programs in education markets may generate spillover effects on students in other schools through competition or other channels. In this case lotteries can measure internally-valid partial equilibrium impacts on applicants but may miss broader general equilibrium effects. Identifying such spillover effects generally requires alternative research designs derived from variation in market structure rather than variation in enrollment opportunities at the student level. Examples of studies in this mold include Figlio and Hart (2014), Gilraine et al. (2021), and Campos and Kearns (2022).

4 Centralized Assignment

4.1 Deferred Acceptance with Single Tie-breaking

Many urban school districts implement district-wide choice using algorithms for centralized assignment.[18] Like the decentralized school admissions lotteries discussed above, many centralized assignment systems incorporate an element of randomness to break ties between students with otherwise identical match criteria. Unlike simple school lotteries, however, the nature of the underlying risk sets in a centralized match is typically shrouded by a seemingly elaborate iterative process. Abdulkadiroğlu et al. (2017) and Abdulkadiroğlu et al. (2021) show how to isolate the random variation in centralized assignment algorithms and how to use this variation to estimate causal effects.

The celebrated Gale and Shapley (1962) deferred acceptance (DA) algorithm is the most widely used for school assignment.[19] To sketch this mechanism, consider a set of N applicants (indexed by i) applying to a set of J schools (indexed by j) with fixed capacities. Applicants submit rank-ordered lists of preferences over schools, defining a set of partial preference orderings denoted by $\succ_i$. Applicants are also given priorities at each school (e.g. those with an enrolled sibling may be highest priority, for instance, followed by applicants who live nearby), denoted by $\phi_{ij} \in \{1, \ldots, P, \infty\}$ where $\phi_{ij} < \phi_{kj}$ means school j prioritizes applicant i over applicant k. The applicant's *type* is defined as $\theta_i = (\succ_i, \phi_i)$, where $\phi_i = (\phi_{i1}, \ldots, \phi_{iJ})$ collects her priorities over all schools; types are collected in $\theta = (\theta_1, \ldots, \theta_N)$. Since priorities are coarse (i.e. there are fewer priority categories than students) student types are further augmented with a set of random tie-breaking numbers $g = (g_1, \ldots, g_N)$ with $g_i \mid \theta \sim U(0, 1)$. Each student's augmented priority is given by $\tilde{\phi}_{ij} = \phi_{ij} + g_i$. The DA mechanism takes as inputs (g, θ) and computes student assignments using the following algorithm:

- Step 0: Each applicant applies to her most preferred school according to $\succ_i$. Each school ranks these applicants by the augmented priority $\tilde{\phi}_{ij}$ and provisionally admits the highest-ranked applicants up to its capacity. All other applicants are rejected.

[18] Cities with centralized school assignment systems include Baltimore, Boston, Cambridge Massachusetts, Camden New Jersey, Chicago, Denver, Indianapolis, Minneapolis, Newark, New York City, New Orleans, Oakland, San Francisco, Seattle, Tulsa, and Washington D.C. Centralized assignment is also widespread and growing globally, with 51 countries using it at either the primary or secondary level as of 2020 (see https://www.ccas-project.org/).

[19] The economic field of *market design* encompasses the study and use of matching tools like Gale-Shapley. Abdulkadiroğlu and Andersson (2022) reviews the market design approach to school choice.

- Step $k > 0$: Each applicant rejected in step $k-1$ applies to her next most preferred school. Each school ranks (by $\tilde{\phi}_{ij}$) these new applicants along with applicants it admitted in step $k-1$. From this pool each school provisionally admits the highest-ranked applicants up to capacity, rejecting the rest.

DA terminates when there are no new applicants, returning a set of assignments, $Z = (Z_i, \ldots, Z_N)$, where $Z_i = j$ indicates assignment of student i to school j.[20]

At a high level, the DA mechanism with single tie-breaking yields a function $M(\cdot)$ from the set of capacities, student types, and lottery numbers to the set of school assignments: $M(g, \theta) = Z$. As with the simple lottery instruments above, the centralized assignments Z_i are likely affected by the random variation in g: students with lower g are more likely to be assigned to a preferable school, all else equal. But unlike simple lotteries there is now a complex translation of this randomness into assignments, $M(\cdot, \theta)$, which depends on the full vector of non-random student types. Students with higher priorities and/or certain preferences are more likely to be assigned to certain schools, regardless of of the tie-breaker they draw. Conditional on applicant type, school offers are ignorable: DA's equal-treatment-of-equals property ensures applicants with the same θ_i have identical school assignment probabilities. As Abdulkadiroğlu et al. (2017) show, however, in large districts such conditioning is impractical since there are almost as many types as students. DA in a high-dimensional scenario generates little useful variation conditional on type.

The Abdulkadiroğlu et al. (2017) solution to this problem leverages the Rosenbaum and Rubin (1983) *propensity score*, defined as the strata-conditional probability of treatment in a stratified RCT. In a DA match with lottery tie-breaking, treatments are indicated by dummies $Z_{ij} = 1\{Z_i = j\}$. The relevant propensity scores are the set of probabilities $p_{ij} \equiv E(Z_i \mid \theta)$, each a scalar function of the high-dimensional list of student preferences and priorities. The Rosenbaum and Rubin (1983) propensity-score theorem implies that in an experiment that randomizes treatment conditional on θ, control for p_{ij} eliminates any OVB arising from the relationship between θ and potential outcomes. In other words, the set of p_{ij} can be used to identify the risk sets induced by centralized assignment.

4.2 Theoretical and Simulated Propensity Scores

In general, the DA propensity score is an unknown function of type. But Abdulkadiroğlu et al. (2017) derive a large-market approximations to the DA propensity score that's easily computed given data on student preferences and priorities. Specifically, Abdulkadiroğlu et al. (2017) derive formulas for centralized assignment propensity scores in a *continuum economy* with a unit mass of students applying to a finite number of schools. Scores for the continuum economy, which are easily computed, approximate finite-market scores remarkably well, and are typically accurate enough to support using Z_{ij} to estimate causal effects.

The large-market approximation works by defining school-specific cutoffs, defined as the last lottery number seated at each school. In the continuum economy, cutoffs are non-random, so each applicant's

[20] The DA assignment is stable in the sense of there being no pair of matched students and schools which would prefer to swap assignments (a "blocking pair"). When students can rank all schools DA is also strategy-proof, in that students have nothing to gain from misreporting their preferences. See Roth and Sotomayor (1990) for a review of these and related concepts.

assignment rate is determined solely by the relevant cutoffs (determined by his or her priorities) and by tie-breaker variation around cutoffs. Applicant assignments conditional on type are independent of one another, a further simplification. The resulting large-market DA propensity score partitions applicants into three groups at each school: applicants who are always, never, and conditionally seated at the school, depending on where their priority for seats falls relative to the school's priority cutoff.[21]

The main payoff to the large-market score is dimension reduction: even in a match with thousands of types, large-market scores are determined by the (relatively) coarse set of school cutoffs. The large-market score also has the side-benefit of distinguishing different sources of centralized assignment risk. Causal effects can, for instance, be estimated separately for conditionally seated applicants and for always seated applicants. Differences in causal effects for these groups can sometimes be linked to economic models of school choice, such as Roy-style selection-on-gains. Abdulkadiroğlu et al. (2017) show that the formulas identifying these different populations apply to other centralized mechanisms, such as random serial dictatorship, and can be extended to DA with multiple tie-breaking (i.e. different lottery numbers at different schools) and the immediate acceptance mechanism (sometimes known as the "Boston mechanism"). Propensity scores for a larger class of stochastic mechanisms satisfying the equal-treatment-of-equals (ETE) property can be simulated by re-drawing the random tie-breaking numbers many times and computing, separately for each applicant, the share of simulations in which the applicant is assigned to a given school. These simulated scores obviate the need for a large-market approximation.

Some centralized assignment schemes, such as those used for Boston and New York City exam schools and New York City screened schools, employ non-lottery tie-breakers such as test scores instead of, or alongside, lottery numbers. Abdulkadiroğlu et al. (2021) show how non-random screening can be combined with lottery variation in a unified *local DA score* approach. Local scores are again derived in a large-market model, with a continuum of applicants and a set of continuously distributed tie-breakers. As before, the large-market model allows a partition of student types into those who are always, never, and conditionally seated and yields simple, coarse formulas for assignment risk.[22] This framework generalizes RD-style identification strategies to settings with multiple treatments and running variables.

Further extensions of the Abdulkadiroğlu et al. (2017) approach to causal inference in centralized assignment systems come from Borusyak and Hull (2020), who show how the propensity score solution can apply to any variable $Z_i = M_i(g, w)$ which combines exogenous shocks g (e.g. random tie-breakers) and non-random data w (e.g. applicant preferences and priorities) according to known formulas $M_i(\cdot)$ (e.g. the DA mechanism).[23] As we discuss further below, controlling for $\mu_i = E[M_i(g, w) \mid w]$, which averages Z_i over

[21] For never seated applicants $p_{ij} = 0$, while for always seated applicants p_{ij} equals the probability that i is not assigned a school she prefers to j. For conditionally seated applicants, p_{ij} is the probability i clears the cutoff at j and does no better than j. Abdulkadiroğlu et al. (2017) show how the latter two probabilities, and thus p_{ij}, are determined by large-market cutoffs for the set of schools in i's rank-order list.

[22] See Section 4.2 of Abdulkadiroğlu et al. (2021) for precise definitions and formulas.

[23] Borusyak and Hull (2020) also discuss local solutions, in which g is viewed as random within a user-specified bandwidth. Note that unlike in the motivating DA mechanism the exogenous shocks in g need not be defined at the same "level" of observations for the general Borusyak and Hull (2020) solution.

the exogenous shocks holding other components fixed, is sufficient to eliminate selection bias in comparisons of individuals assigned different values of Z_i. This generalizes the propensity score approach to settings with multi-valued or continuous treatments. Importantly, this result holds even when Z_i is not generated from a mechanism satisfying the ETE property, or indeed outside of the case where Z_i indicates a centralized school assignment. Borusyak and Hull (2020) compute μ_i by repeatedly drawing g and averaging the resulting Z_i over draws, holding fixed the non-random w. While potentially computationally demanding, this simulation procedure yields a general recipe for extracting useful variation from a complex (but known) assignment scheme.

DA also generates a variety of other ad hoc instruments, with simpler propensity scores. One is a dummy for whether an applicant is offered a seat at their first-choice school. This first-choice assignment dummy is randomly assigned conditional on first-choice preference and priority risk sets, which may be simply controlled for. Alternatively, qualification instruments indicate whether g_i is better than the worst lottery number offered a seat (controlling for the set of schools ranked). Although valid, first choice and qualification instruments are likely to leave much useful assignment variation on the table (Narita, 2016); we illustrate this phenomenon below.[24]

4.3 Estimation with Score Controls

Once computed, centralized assignment propensity scores can be used in a variety of ways to estimate school effectiveness. Abadie (2003), for example, proposes estimators for LATEs and related parameters that inversely weight by instrument propensity scores. Propensity score matching, as proposed by Rosenbaum and Rubin (1983), is another option for obtaining an equal-weighted average of local causal effects. A simple alternative is to adjust for assignment propensity scores in a linear IV regression. For a given school j, consider the IV second and first stages of

$$Y_i = \beta D_i + X_i'\mu + \eta_i, \tag{19}$$

$$D_i = \pi \tilde{Z}_{ij} + X_i'\psi + \upsilon_i, \tag{20}$$

where here D_i indicates enrollment in some school j, X_i is a vector of predetermined controls, and $\tilde{Z}_i = Z_{ij} - p_{ij}$ is a "recentered" offer instrument that subtracts the propensity score p_{ij} from the offer indicator Z_{ij}. Borusyak and Hull (2020) show how such specifications identify weighted averages of conditional-on-type IV coefficients. In particular, the IV estimand is given by:

$$\beta_{IV} = \int w_{IV}(t)\beta_{IV}(t)dF_\theta(t), \tag{21}$$

[24] Examples of the first choice IV in centralized assignment mechanisms include Deming (2011), Abdulkadiroğlu et al. (2013), Deming et al. (2014), and Hastings et al. (2009). Examples of the qualification IV include Dobbie and Fryer (2014), Lucas and Mbiti (2014), and Pop-Eleches and Urquiola (2013). First-choice instruments have also been used with decentralized assignment mechanisms (Abdulkadiroğlu et al., 2011; Cullen et al., 2006; Dobbie and Fryer, 2011; Hoxby et al., 2009).

where $F_\theta(\cdot)$ gives the distribution of types θ_i and

$$\beta_{IV}(t) = \frac{E[Y_i|Z_{ij}=1,\theta_i=t] - E[Y_i|Z_{ij}=0,\theta_i=t]}{E[D_i|Z_{ij}=1,\theta_i=t] - E[D_i|Z_{ij}=0,\theta_i=t]}$$

is the Wald estimand for students of type $\theta_i = t$. Analogous to the weighting function in (8), the weights $w_{IV}(t)$ in (21) integrate to 1 and are proportional to the share of students of type $\theta_i = t$, the conditional-on-type assignment variance $Var(Z_{ij} \mid \theta_i = t)$, and the conditional first stage $E[D_i|Z_{ij}=1,\theta_i=t] - E[D_i|Z_{ij}=0,\theta_i=t]$. The IV coefficient β_{IV} in (19) thus equals a convex average of the $\beta_{IV}(t)$ when assignment weakly increases enrollment, and can be interpreted as a weighted-average of LATEs when conditional-on-type analogs of Assumptions 2A and 2C hold.

Equations (19) and (20) can be seen as generalizing the risk-set-controlled IV specification (6)-(7) for a high-dimensional type vector θ_i that is linearly adjusted for via the propensity score p_{ij} rather than though a set of risk set indicators. The connection is strengthened by noting the estimand is unchanged when p_{ij} is included in the control vector X_i and the recentered instrument $\tilde{Z}_i$ is replaced with the unadjusted offer Z_{ij}. This observation follows from the Frisch-Waugh-Lovell theorem: the residual from regressing Z_{ij} on p_{ij} and X_{ij} is $\tilde{Z}_i = Z_{ij} - p_{ij}$, since $p_{ij} = E[Z_{ij} \mid \theta_i]$ predicts Z_{ij} with a coefficient of one and all other predetermined variables in X_i are independent of Z_{ij} controlling for p_{ij}. Thus we obtain the same IV estimand shown in (21) by recentering Z_{ij} by p_{ij} or by controlling for a linear function of p_{ij}, or any more flexible function of p_{ij} in 2SLS estimation (with or without other predetermined covariates X_i).[25]

As with equation (8), linear adjustment for p_{ij} is likely to yield precise estimates of school j's effectiveness by efficiently aggregating all conditionally random variation in centralized assignments. The weights $w_{IV}(t)$ discard types which are always or never assigned to j regardless of the random tie-breaker, while putting more weight on types where assignment and non-assignment are equally likely. Further precision gains can be achieved by adding predetermined controls that predict residual outcome variation. Controlling for p_{ij} instead of recentering Z_{ij} can thus lead to smaller standard errors in practice. Abdulkadiroğlu et al. (2017) go a step further by controlling for indicators for each value of the propensity score as well as student demographics and lagged achievement measures.

Centralized school offers and propensity scores can serve as the building blocks for empirical examinations of the effectiveness of school sectors (such as charters) and the effect of assignment to schools with different characteristics (such as those with high district ratings or with certain peer characteristics). Formally, let C_j denote a characteristic of school j, and let $z(i)$ and $d(i)$ denote the indices of student i's assigned and enrolled schools. Consider the instrument $C_{z(i)} = \sum_j C_j Z_{ij}$ and treatment $C_{d(i)} = \sum_j C_j D_{ij}$ measuring, respectively, the characteristic of the assigned and enrolled schools. With C_j indicating charter schools, for example, $C_{z(i)}$ is an indicator for being assigned to a charter while $C_{d(i)}$ indicates charter enrollment. The

[25] An additional application of the Frisch-Waugh-Lovell theorem shows that 2SLS estimation of the risk-set-controlled model (6)-(7) is equivalent to controlling for the empirical propensity score calculated as the mean offer rate within each risk set. Control for the theoretical score generated by the assignment mechanism makes estimation feasible when nonparametrically estimating the score for each risk set is not feasible.

Borusyak and Hull (2020) characterization extends to IV estimators which instrument $C_{d(i)}$ with $C_{z(i)}$ while controlling for $\sum_j C_j p_{ij}$: an average of school characteristics weighted by the assignment propensity scores.[26] For examining charter school effectiveness this would mean controlling for the total risk of assignment to charters. We return to other school characteristic IVs in the next section.

Table 6 illustrates the role played by centralized assignment propensity scores through an analysis of charter effects in Denver Public Schools (the sample used here is from Abdulkadiroğlu et al. (2017)). Column (1) shows a precise 0.42σ charter attendance effect on math test scores, estimated by 2SLS with a charter offer instrument controlling flexibly for the simulated charter propensity score and other baseline demographics.[27] The remaining columns show similar but less precise estimates obtained from cruder instruments: an indicator for first-choice charter assignment in column (2), and an indicator for qualification for any-charter assignment in column (3), both controlling for for the appropriate preference-based risk sets. These alternative strategies discard some of the random variation in charter assignment generated by the mechanism, and therefore produce less precise estimates, as reflected in the second-stage standard errors. The second-to-last row of Table 6 demonstrates that sample sizes for the first-choice and qualification approaches would need to increase by factors of 1.6 and 3.5 to match the precision of the risk-controlled centralized assignment IV strategy.

5 VAM for Individual Schools

For many high-stakes decisions, knowing the effectiveness of a broad school sector (such as charters vs. traditional public schools) is insufficient. Parents wish to know which schools, in particular, deliver the most learning for their children. Policymakers likewise may rely on individual school effectiveness measures when deciding whether to close, restructure, or expand schools in their district (Rockoff and Turner, 2010; Abdulkadiroğlu et al., 2016; Cohodes et al., 2021). This demand for school effectiveness data is reflected in a recent proliferation of publicly available measures. The 2015 Every Student Succeeds Act, for example, mandated all US states to adopt elementary and middle school accountability systems that include public measures of average student achievement and growth. Private companies such as US News and World Report and GreatSchools.org, meanwhile, produce massively popular school ratings that are often featured prominently on real estate sites like Zillow and Redfin. Such ratings appear to affect families' choices of where to live, as well as where to enroll their children (Bergman and Hill, 2018; Hasan and Kumar, 2019).

Virtually all public and commercial school performance measures are derived from observational comparisons: typically, average test score levels or growth among a school's enrolled students, sometimes adjusted for differences in observed student demographics. Such levels and growth measures closely resemble the observational value-added models (VAMs) that have long been considered and debated for ranking teachers

[26] As before, this specification identifies the same coefficient as the IV procedure which instruments with the recentered $\sum_j C_j(Z_{ij} - p_{ij})$. Predetermined controls can be included in either regression to increase precision.

[27] The sample includes applicants for grades 4-10 in the 2011-2012 and 2012-2013 school years. See Tables 6 and 9 of Abdulkadiroğlu et al. (2017) for more details.

and schools (e.g. Kane and Staiger, 2008; Rothstein, 2010; Chetty et al., 2014a; Deming, 2014).[28] The selection-on-observables assumption which underlies such VAMs reflects a different approach to removing selection bias than the lottery and discontinuity-based identification strategies from above. Yet a recent literature, starting with Angrist et al. (2016b, 2017), shows how such quasi-experimental variation can be incorporated in the school VAM agenda, yielding more reliable estimates of individual school effectiveness while grappling with certain fundamental issues with obtaining such fine-grained causal estimates.

In this section we first overview the basic logic of observational school VAMs and empirical Bayes methods that are commonly applied to value-added estimates. Next, we discuss how the key identifying assumptions of observational VAMs can be tested with school lotteries. We then describe ways such variation can be further used to improve observational models by partially correcting for selection bias when the identifying assumptions are found to be violated.

5.1 Estimating Observational VAMs

The starting point for conventional school value-added estimation is a constant-effects model along the lines of those considered earlier, now extended to allow each school to have a distinct causal effect. Consider a setting with multiple schools indexed by $j = 1, \ldots, J$, each with its own causal "value-added" parameter β_j measuring its effect relative to an omitted school on some achievement outcome Y_i:

$$Y_i = \mu + \sum_{j=1}^{J} \beta_j D_{ij} + \varepsilon_i. \tag{22}$$

As before $D_{ij} \in \{0, 1\}$ indexes the enrollment of student i in school j and ε_i captures other determinants of achievement.[29]

To estimate the value-added parameters β_j, policymakers and practitioners may use a combination of regression controls and outcome-differencing strategies. Regressing the unobserved ε_i on a vector of observed covariates X_i, which again may include student demographics and lagged test scores, yields an augmented causal model:

$$Y_i = \sum_{j=1}^{J} \beta_j D_{ij} + X_i'\mu + \eta_i, \tag{23}$$

where (also as before) in the transition to a model with covariates we repurpose the symbol μ to be the coefficient vector attached to X_i, which is defined to include a constant. The key selection-on-observables assumption underlying VAM estimation is that $E[D_{ij}\eta_i] = 0$ for each j. In other words, the component of

[28] Observational examinations of school effectiveness can be traced at least as far back as to the Coleman (1966) report, which famously showed in cross-sectional regressions that the fraction of variance in student achievement attributable to educational inputs was small relative to the contribution of family background. Suffice to say the long observational and quasi-experimental literatures that followed paint a more nuanced picture.

[29] Formally, equation (22) derives from an additively-separable potential outcomes model, $Y_i(j) = \mu + \beta_j + \varepsilon_i$, which implies causal effects $Y_i(j) - Y_i(k) = \beta_j - \beta_k$ are constant across students.

student ability that is unexplained by X_i must be uncorrelated with school enrollment. Under selection-on-observables, an OLS regression recovers the parameters of equation (23).

This framework nests several types of school quality measures. The simplest *levels* measures effectively set X_i equal to only a constant, thereby measuring school quality as the average achievement level of enrolled students. Estimates from levels models will be biased when schools enroll students with systematically different unadjusted ability ε_i, as seems likely since schools are not randomly assigned. More sophisticated VAMs account for observable differences in demographics and lagged test scores by including these controls in X_i. In this case the selection-on-observables assumption requires that there is no systematic selection into school enrollment among students with the same characteristics and past achievement.

An alternative approach to adjusting for lagged achievement is a *gains* model that first-differences contemporaneous and past test scores to remove time-invariant unobservables. If we let $\Delta Y_i = Y_{ig} - Y_{i(g-1)}$ denote the change in student achievement relative to an earlier grade, the gains model is given by:

$$\Delta Y_i = \mu + \sum_{j=1}^{J} \beta_j D_{ij} + \Delta\varepsilon_i, \tag{24}$$

where $\Delta\varepsilon_i = \varepsilon_i - Y_{i(g-1)}$. Here the relevant identification assumption requires that potential outcome trends are parallel across schools, i.e. $E[\Delta\varepsilon_i \mid D_{ij} = 1] = E[\Delta\varepsilon_i \mid D_{ik} = 1]$ for all $j \neq k$, so that a linear regression of ΔY_i on the D_{ij} recovers the value-added parameters β_j. The regression adjustment and outcome differencing in equations (23)-(24) can further be combined by adding additional demographic controls to the gains regression.

5.2 Empirical Bayes Methods

EB Shrinkage Under Normality

OLS estimation of VAM models like (23) and (24) yields a set of school-specific value-added estimates $\hat{\beta}_j$.[30] One key question, which we turn to in the next subsection, is whether the identifying assumptions underlying these observational VAM procedures hold such that $\hat{\beta}_j$ gives an unbiased estimate of the causal parameter β_j from equation (22). Setting aside this question for the moment, another perhaps equally important one is whether $\hat{\beta}_j$ is estimated precisely enough to be useful for decision-making even when these identifying assumptions hold. Compared to the sector-wide effect estimates discussed in Section 3, single-school VAM estimates are likely to involve substantial sampling error, particularly for small or new schools with relatively few student observations.

Empirical Bayes (EB) analysis offers a strategy to moderate sampling variance in individual estimates of

[30] In some applications researchers instead estimate value-added by first regressing Y_i on X_i, then computing school averages of the resulting residuals. This approach generates the same estimates as OLS estimation of (23) in large samples if school enrollment is independent of the controls X_i, but generally yields asymptotically different estimates if enrollment is correlated with the controls. Since the inclusion of X_i is typically motivated by concerns about selection bias, it seems preferable when possible to use the OLS estimates which do not impose this independence assumption.

$\hat{\beta}_j$. The EB approach treats the school VAM parameters β_j as draws from a distribution of school quality, typically assumed to be Normal. EB estimates of the *hyperparameters* that characterize this distribution are derived from the estimated $\hat{\beta}_j$'s. This estimated distribution generates posterior predictions for individual school quality. The EB approach uses the full set of VAM estimates to reduce sampling variance in individual school quality estimates, accepting bias in the posterior estimates in exchange (Morris, 1983; Raudenbush and Bryk, 1986; Efron, 2012).

By way of illustration, suppose $\hat{\beta}_j$ is an unbiased and Normally-distributed estimate of β_j with known variance equal to its squared standard error s_j^2. This Normality assumption can be viewed as an asymptotic approximation with a growing number of students per school. Next, suppose the latent parameters β_j are themselves drawn randomly from a distribution G_β defined in the population of schools. Assume for the moment that G_β is a Normal distribution and independent of sampling variance s_j^2 across schools. This yields the following hierarchical model:

$$\hat{\beta}_j|\beta_j, s_j^2 \sim N(\beta_j, s_j^2), \tag{25}$$

$$\beta_j|s_j^2 \sim N(\mu_\beta, \sigma_\beta^2). \tag{26}$$

This model has two hyperparameters, μ_β and σ_β^2.[31] Method of moments estimates of these hyperparameters are given by

$$\hat{\mu}_\beta = \frac{1}{J}\sum_{j=1}^{J}\hat{\beta}_j, \tag{27}$$

$$\hat{\sigma}_\beta^2 = \frac{1}{J}\sum_{j=1}^{J}[(\hat{\beta}_j - \hat{\mu}_\beta)^2 - s_j^2]. \tag{28}$$

The variance estimator here subtracts s_j^2 in (28) as a bias correction: the uncorrected sample variance of $\hat{\beta}_j$'s is inflated by sampling variance.[32] Maximum likelihood applied to the school-level model (25) and (26), or a full parametric specification starting with distributional assumptions for the residual in model (23) for individual outcomes, offer more elaborate alternatives to the simple method of moments approach.

The final step in EB estimation constructs posteriors for the quality of each school. Given the model in (25) and (26), the posterior distribution for β_j is given by $\beta_j|\hat{\beta}_j, s_j^2 \sim N(\beta_j^*, V_j^*)$ where

$$\beta_j^* = \left(\frac{\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}\right)\hat{\beta}_j + \left(\frac{s_j^2}{\sigma_\beta^2 + s_j^2}\right)\mu_\beta, \tag{29}$$

and $V_j^* = \frac{s_j^2\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}$. Equation (29) shows that the posterior mean β_j^* is a weighted average of the unbiased estimate $\hat{\beta}_j$ and the prior mean μ_β. The weight on $\hat{\beta}_j$ approaches one as its sampling variance s_j^2 approaches

[31] Given the model in (22), the mean μ_β captures mean school quality relative to an omitted category.

[32] Kline et al. (2020) outline general methods for bias-corrected estimation of variance components.

zero. By shrinking the noisy unbiased estimate toward the prior mean in proportion to its sampling error, the posterior mean reduces variance, with more shrinkage for schools with noisier estimates. An empirical Bayes posterior mean $\hat{\beta}_j^*$ plugs the estimated hyperparameters $\hat{\mu}_\beta$ and $\hat{\sigma}_\beta^2$ into (29).

When to Shrink?

Whether the shrunk EB posterior mean should be preferred to the noisier but unbiased estimate $\hat{\beta}_j$ depends on the goals of the analyst. To see this, note that conditional on the true value-added of school j, the mean squared error (MSE) of the two estimates is:[33]

$$\begin{aligned} &E[(\hat{\beta}_j - \beta_j)^2 | \beta_j, s_j^2] = s_j^2, \\ &E[(\beta_j^* - \beta_j)^2 | \beta_j, s_j^2] = \left(\frac{\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}\right)^2 s_j^2 + \left(\frac{s_j^2}{\sigma_\beta^2 + s_j^2}\right)^2 (\beta_j - \mu_\beta)^2 \,. \end{aligned} \tag{30}$$

If we are only interested in one specific school, this conditional MSE formula shows it's not clear which of the two estimators is better; shrinkage reduces variance but may lead to substantial bias if the school is very different from average (as reflected in the second term in equation (30)). On the other hand, if we are interested in evaluating many schools, the relevant notion of MSE integrates over the distribution of β_j:

$$\begin{aligned} &E[(\hat{\beta}_j - \beta_j)^2 | s_j^2] = s_j^2, \\ &E[(\beta_j^* - \beta_j)^2 | s_j^2] = \left(\frac{\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}\right) s_j^2. \end{aligned} \tag{31}$$

This formula shows that in an unconditional sense the posterior mean has unambiguously lower MSE than the unbiased estimate $\hat{\beta}_j$.[34] In fact, by standard properties of conditional mean functions, the posterior mean has lowest MSE of all functions of $(\hat{\beta}_j, s_j^2)$ under the model. EB methods are therefore useful when we want an estimator that performs well on average across all schools.

Once EB posterior distributions have been calculated they can be used for several purposes.[35] First, as shown in equation (31), EB shrinkage yields a set of estimates with low average MSE across schools. Second, shrinkage corrects measurement error in models that treat school value-added as a regressor. Putting the unbiased but noisy estimate $\hat{\beta}_j$ on the right-hand side of a regression results in attenuation bias towards zero due to classical measurement error; the posterior mean introduces non-classical measurement error that corrects this, so that a regression with β_j^* on the right yields the same coefficient as using the true β_j.

[33] The MSE formula for the posterior mean ignores estimation error in the hyperparameters. See Morris (1983) and Armstrong et al. (2020) for discussion of approaches that incorporate estimation error in the prior distribution.

[34] This observation is closely related to the classic James and Stein (1961) result that OLS is inadmissible and dominated by shrinkage-based estimators when estimating three or more parameters under quadratic loss.

[35] Analysts sometimes report the variance of posterior mean estimates as a summary of the dispersion in value-added. Such a procedure can yield misleading results because the distribution of posterior means is by construction underdispersed relative to the underlying distribution of latent parameters (formally, $Var(\beta_j^*) < \sigma_\beta^2 < Var(\hat{\beta}_j)$). For the purpose of understanding the variance of value-added, the hyperparameter estimate $\hat{\sigma}_\beta^2$ is more useful.

This correction is closely related to traditional errors-in-variables regression, which similarly corrects for measurement error using an estimated signal-to-noise ratio but typically uses a common shrinkage factor for all observations (Draper and Smith, 1998).[36]

Third, EB posterior distributions can be used for making decisions about individual schools. For this purpose, the right feature of the posterior to use will depend on the decision-maker's loss function, and in general it may be optimal to use functionals other than the posterior mean. For example, a district policymaker might be interested in closing all schools below a quality threshold $\bar{\beta}$ and view each mistaken closure as equally costly, in which case it is optimal to make decisions based on the posterior probability that β_j is less than $\bar{\beta}$ (given by $\Phi\left(\frac{\bar{\beta}-\beta_j^*}{\sqrt{V_j^*}}\right)$ in the parametric Normal/Normal model). Recent work by Gu and Koenker (2020) discusses EB methods for ranking and tail selection decisions.

EB Extensions

The simple EB framework sketched here is usefully extended by adding covariates that predict school quality and by allowing a more flexible form for the prior distribution. It's useful in some cases to have $\beta_j | s_j^2, C_j \sim N(C_j'\mu, \sigma_\beta^2)$ given a list of school characteristics, C_j, such as school sectors. The resulting EB posterior mean shrinks $\hat{\beta}_j$ toward an estimated linear index $C_j'\hat{\mu}$ (estimable from a regression of $\hat{\beta}_j$ on C_j) rather than a constant. It is straightforward to also allow σ_β^2 to depend on school characteristics.

In the same spirit, Normality of (26) might be replaced by a more general model for G_β. The linear shrinkage estimator derived under Normality has desirable properties regardless. In particular, β_j^* coincides with the fitted value from a linear regression of β_j on $\hat{\beta}_j$, so it can be interpreted as a best linear predictor of value-added regardless of the mixing distribution. It may nonetheless be of interest to estimate a more flexible G_β to obtain a more complete picture of the shape of the school quality distribution and form improved posteriors. Methods for this purpose include the Kiefer and Wolfowitz (1956) non-parametric maximum likelihood estimator (NPMLE; see also Robbins (1956) and Koenker and Mizera (2014)) and the exponential family deconvolution estimator proposed by Efron (2016). Gilraine et al. (2020) apply the NPMLE approach to estimation of teacher value-added distributions.[37]

In some other cases, it's useful to allow β_j and s_j^2 to be correlated. For example, it could be that newer or smaller schools are less effective, in which case schools with more students (and therefore lower s_j^2) would tend to have higher β_j. A simple strategy here treats s_j^2 as a covariate in the conditional distribution of G_β; the NPMLE estimator can also be applied to estimate an unrestricted bivariate distribution of β_j and s_j^2. An alternative approach is to apply a variance-stabilizing transform to the $\hat{\beta}_j$'s that results in approximately constant sampling variance before estimating the prior distribution (see, e.g., Brown (2008)). Since the empirical literature on school quality using EB methods to date has typically relied on Normality

[36] Note that since classical measurement error on the left does not cause bias, putting β_j^* rather than $\hat{\beta}_j$ on the left-hand side of a regression causes bias rather than correcting it.

[37] See Kline and Walters (2021) and Kline et al. (2021) for other recent applications of non-parametric EB methods outside of education.

and independence assumptions, the practical relevance of relaxing these assumptions is unclear.

5.3 Testing VAM Validity with Lotteries

Of course, a precise but highly biased estimate of posterior of school value-added is likely to be as problematic as an extremely noisy estimate. We next consider the use of quasi-experimental school assignment variation to test for selection bias in observational VAMs. The key assumption of an observational VAM is selection-on-observables: conditional on a set of included controls X_i, school enrollment is as good as randomly assigned. This assumption can be tested with the help of a regression of the form:

$$Y_i = \sum_j \alpha_j D_{ij} + X_i'\mu + \nu_i, \tag{32}$$

where X_i is the vector of controls in the causal model, (23). This regression model differs from (23) only in that it is *defined* as a regression and so may have parameters different from those in the causal model. The error term in this model, ν_i, therefore differs from the random part of potential outcomes, η_i.

Under selection-on-observables, the parameters of the causal model (23) and the regression (32) coincide so that $\alpha_j = \beta_j$ for all schools j. Moreover, the regression residual coincides with residual ability: $\nu_i = \eta_i$. Student ability should, in turn, be unrelated to any randomness in school offers for the same reasons underlying Assumption 1A. Thus, a consequence of selection on observables is orthogonality of OLS residuals and (recentered) assignment indicators $Z_{i\ell}$:

$$E[(Z_{i\ell} - p_{i\ell})\nu_i] = 0, \tag{33}$$

for potentially multiple schools $\ell = 1, \ldots, L$. Here we are using the assignment propensity scores $p_{i\ell}$ to address any non-randomness in offers, as in the centralized assignment scenario of Section 4, but the same logic can be applied to decentralized lotteries by adjusting for risk set fixed effects, as in Sections 2-3.

A test of (33) is obtained by regressing the residuals from regression model (32) on the set of instruments, $Z_{i\ell} - p_{i\ell}$. Test rejections are symptomatic of selection bias in the observational VAM coefficients: i.e. that $\nu_i \neq \eta_i$ such that $\alpha_j \neq \beta_j$ for some or all schools j. Angrist et al. (2016b) show how this procedure can be viewed as a Lagrange multiplier (LM) test of the L orthogonality restrictions, which impose the joint null hypothesis of VAM validity and conditionally independent school assignment.

An omnibus test of (33) can be decomposed into two conceptually distinct tests: one capturing the extent to which the observational VAM coefficients α_j predict school effectiveness β_j on average, and one capturing how any selection bias $b_j \equiv \alpha_j - \beta_j$ varies across schools. Formally, consider a test statistic constructed based on an assumption of homoskedastic ν_i:

$$\hat{T} = \frac{(Y - \hat{Y})' P_Z (Y - \hat{Y})}{\hat{\sigma}_\nu^2},$$

where Y is an $N \times 1$ vector of the achievement outcome, $\hat{Y}$ is an $N \times 1$ vector of regression fitted values from estimating equation (32), $\hat{\sigma}_\nu^2 = (Y - \hat{Y})'(Y - \hat{Y})/N$ estimates the variance of ν_i, and P_Z is the projection matrix for the recentered offers $Z_{i\ell} - p_{i\ell}$.[38] This is a joint test of significance of the regression of VAM residuals $\hat{\nu}_i = Y_i - \hat{Y}_i$ on $Z_{i\ell} - p_{i\ell}$:

$$\hat{\nu}_i = \psi_0 + \sum_\ell \psi_\ell (Z_{i\ell} - p_{i\ell}) + u_i,$$

where the null of $\psi_1 \dots, \psi_L = 0$ is imposed to compute the residual variance. Angrist et al. (2016b) show this test statistic can be rewritten as the sum of two terms:

$$\hat{T} = \frac{(\hat{\varphi} - 1)^2}{\hat{\sigma}_\nu^2 (\hat{Y}' P_Z \hat{Y})^{-1}} + \frac{(Y - \hat{\varphi}\hat{Y})' P_Z (Y - \hat{\varphi}\hat{Y})}{\hat{\sigma}_\nu^2}, \tag{34}$$

where $\hat{\varphi} = (\hat{Y}' P_Z \hat{Y})^{-1} \hat{Y}' P_Z Y$ is the 2SLS coefficient estimate that uses all $Z_{i\ell} - p_{i\ell}$ to instrument $\hat{Y}_i$ in an equation for Y_i. Since these recentered offers are by design uncorrelated with student demographics or lagged test scores, $\hat{\varphi}$ can be seen to estimate the second-stage equation

$$Y_i = \tau_0 + \varphi \alpha_{d(i)} + X_i' \tau + \eta_i, \tag{35}$$

where $\alpha_d(i) = \sum_j \alpha_j D_{ij}$ denotes the observational VAM coefficient of i's enrolled school.

The second-stage parameter φ is sometimes referred to as a "forecast coefficient," when versions of equation (35), fit with quasi-experimental variation, are used to assess the on-average predictive validity of observational quality measures α_j—whether for teachers, schools, or more recently outside of education (e.g. Chetty et al., 2014a; Deming, 2014; Chetty and Hendren, 2018; Abaluck et al., 2021). The typical null hypothesis is that $\varphi = 1$, meaning that a one-unit increase in $\alpha_{d(i)}$ translates into a one-unit increase in Y_i. Deviations from this predictive relationship indicate "forecast bias." The first term of the Angrist et al. (2016b) test statistic decomposition (34) is therefore a Wald statistic for the null of no forecast bias. The second term is the Sargan (1958) statistic for an LM test of the overidentifying restrictions in 2SLS estimation of equation (35). Intuitively, this term checks whether the VAM coefficients are equally predictive within each lottery quasi-experiment. The omnibus test of VAM validity $\hat{T}$ checking the L restrictions (33) combines a single test of forecast bias with $L - 1$ additional restrictions coming from various school-specific admissions lotteries.

Table 7 illustrates these tests for Boston middle schools using the sample of students and lotteries analyzed in Angrist et al. (2017). The first row reports forecast coefficients from 2SLS models instrumenting VAM predictions $\hat{\alpha}_{d(i)}$ with charter school lottery offers and first-choice centralized assignments, controlling for the necessary risk set fixed effects and additional baseline covariates. Column (1) shows results for an *uncontrolled* VAM which simply compares average 6th grade math achievement across schools adjusting for

[38] Formally, $P_Z = Z(Z'Z)^{-1}Z'$ where Z is a $N \times L$ matrix stacking observations of the recentered assignments $Z_{i\ell} - p_{i\ell}$.

year effects. The *lagged score* model in column (2) further adjusts for student demographics and lagged (5th grade) achievement, while the *gains* model in column (3) replaces the outcome with the difference in achievement and lagged achievement, maintaining the demographic covariates.

The test results reveal that VAM models adjusting for lagged achievement are much less biased than the naive uncontrolled specification. The 2SLS forecast coefficient estimate for the uncontrolled model is only 0.40, indicating substantial forecast bias in comparisons of unadjusted achievement levels across schools. In contrast, the lagged score specification generates a forecast coefficient of 0.86 that is only marginally statistically different from one ($p = 0.07$). The corresponding estimate for the gains model equals 0.95, and the null hypothesis of no forecast bias in the gains model cannot be rejected at conventional levels ($p = 0.55$). This minimal forecast bias reflects a common finding in the literatures on teacher and school value-added: approaches that adjust for past achievement eliminate much of the selection bias in comparisons across classrooms and schools, so that on average estimates from a well-controlled VAM provide reasonably reliable estimates of causal effects on student test scores (Chetty et al., 2014a; Angrist et al., 2017).

Despite this minimal forecast bias, the final row of the table shows that the omnibus test of all restrictions rejects decisively for all three models. The rejection is driven primarily by failure of the 2SLS overidentifying restrictions: even though the gains model predicts student outcomes well on average, this predictive validity is poor for some of the L school-specific lotteries. Test results for the more sophisticated VAMs in columns (2) and (3) are depicted graphically in Figure 3, which gives a “visual IV” representation of the 2SLS estimates. Specifically, for each of the Angrist et al. (2017) school lotteries we plot the reduced-form effect of assignment on 6th grade math scores Y_i against the first-stage effect of assignment on estimated observational value-added $\hat{\alpha}_{d(i)}$. The slopes of the weighted lines-of-best-fit through these points correspond to the forecast coefficient estimates $\hat{\varphi}$ from the table, and we plot the benchmark 45-degree line for reference. As we've seen, the gains model brings these two lines closer together, reflecting minimal forecast bias. But several lottery points are far from these lines, suggesting non-zero selection bias across schools (colored points are statistically significantly far from this line at the 10% level).

5.4 Bias-correction with Lotteries

Combining OLS and IV Estimates

Given test rejections like in Table 7, a natural question is whether quasi-experimental admissions variation can be used to reduce the apparent selection bias in observational VAMs. In a setting where both lottery-based and observational VAM estimates of school quality are available, there is a tradeoff between using biased (but precise) observational estimates and unbiased (but noisy) lottery estimates. We next sketch a potential solution to this tradeoff by extending the empirical Bayes framework of Section 5.2.

Suppose we have a set of potentially biased observational OLS VAM estimates from equation (23), satisfying:

$$\hat{\alpha}_j|\beta_j, b_j, s^2_{j,\alpha} \sim N(\beta_j + b_j, s^2_{j,\alpha}). \tag{36}$$

As before, parameter β_j gives the true quality of school j, $s^2_{j,\alpha}$ is the squared standard error of the OLS estimate $\hat{\alpha}_j$, and the Normality assumption is an asymptotic approximation with many students per school. Now, however, the estimator may be biased—represented by the school-specific parameter b_j. A rejection of the lottery-based omnibus test in Section 5.3 indicates that $b_j \neq 0$ for some or all schools.

Suppose in addition to the $\hat{\alpha}_j$ estimates we have a quasi-experimental VAM estimate $\hat{\beta}_j$ for each school. For example, these estimates might come from instrumenting the D_{ij} indicators in equation (23) with risk-adjusted offer instruments $Z_{ij} - p_{ij}$ as described in Section 4. The lottery estimates are assumed to be consistent and asymptotically normal estimates of the true VAM parameters:

$$\hat{\beta}_j|\beta_j, b_j, s^2_{j,\beta} \sim N(\beta_j, s^2_{j,\beta}). \tag{37}$$

We generally expect $s^2_{j,\beta}$ to be larger than $s^2_{j,\alpha}$ since IV is less precise than OLS.[39] Finally, we extend the hierarchical model (26) to allow for a bivariate distribution of school quality and selection bias across schools. Letting $\Theta_j \equiv (\beta_j + b_j, \beta_j)'$ denote the 2×1 vector of observational VAM and causal parameters for school j, write:

$$\Theta_j|S_j \sim N(\mu_\Theta, \Sigma_\Theta). \tag{38}$$

The matrix Σ_Θ describes the joint distribution of causal effectiveness and selection bias across schools. The matrix S_j has the sampling variances $s^2_{j,\alpha}$ and $s^2_{j,\beta}$ on the diagonal and the sampling covariance of the OLS and IV estimates off the diagonal. When $b_j = 0$ and the OLS residual ν_i is homoskedastic, this covariance equals $s^2_{j,\alpha}$ (Hausman, 1978).

Under the model described by (36), (37), and (38), the posterior mean for Θ_j conditional on the observed estimates $\hat{\Theta}_j = (\hat{\alpha}_j, \hat{\beta}_j)'$ is given by

$$\Theta^*_j \equiv E[\Theta_j|\hat{\Theta}_j, S_j] = (\Sigma^{-1}_\Theta + S^{-1}_j)^{-1} S^{-1}_j \hat{\Theta}_j + (\Sigma^{-1}_\Theta + S^{-1}_j)^{-1} \Sigma^{-1}_\Theta \mu_\Theta. \tag{39}$$

The second element of Θ^*_j is the posterior mean for β_j using both the OLS and IV estimates of school quality, which is a linear combination of the two estimates and the prior mean.[40]

Following the EB approach of Section 5.2, the "hybrid" value-added posterior means in (39) can be approximated by estimating hyperparameters μ_Θ and Σ_Θ, based on the observed joint distribution of OLS and IV estimates and their sampling variances and covariances, and then plugging these hyperparameter estimates into equation (39). By the above Hausman (1978) logic on S_j, it can be shown that when the observational VAM is close to unbiased ($Var(b_j) \approx 0$) and the errors are homoskedastic no weight is placed

[39] Note that this would be guaranteed under the classical assumptions of the Gauss-Markov model.

[40] See Angrist et al. (2017) for expressions for the weights on the two estimates in special cases. Chetty and Hendren (2018) use a similar approach to compute EB estimates of neighborhood effects combining observational and quasi-experimental variation.

on the IV estimates and we return to the conventional EB shrinkage formula (29) applied to the OLS $\hat{\alpha}_j$ estimates. At the opposite extreme, when selection bias is bad enough to make the observational VAM estimates useless (i.e. $Var(b_j) \to \infty$), no weight is placed on $\hat{\alpha}_j$ and we arrive at a conventional EB shrinkage formula for the quasi-experimental estimates $\hat{\beta}_j$. In intermediate cases the hybrid posterior optimally trades off bias and variance between the two sets of value-added estimates.

Bias-correction with Undersubscription: IV VAM

In practice a lottery-based estimate $\hat{\beta}_j$ may not be available for every school. In centralized assignment systems, for example, there are typically some schools that lack quasi-experimental variation in assignment. This is a problem of *undersubscription*, formalized as $Z_{ij} - p_{ij} = 0$ for all students i at some school j. Applicants who are assigned to such schools ($Z_{ij} = 1$) never face any risk of non-assignment ($p_{ij} = 1$), perhaps because the school faces weak demand and so has no need for lottery-based rationing, while all other students are never assigned ($Z_{ij} = p_{ij} = 0$). With $Z_{ij} - p_{ij} = 0$ for all students, assignment at this school cannot instrument for school enrollment; given undersubscription, therefore, we have fewer instruments than endogenous variables in equation (23) and cannot use 2SLS to estimate the β_j's or apply the hybrid posterior formula (39).

An instrumental variables value-added model (IV VAM) approach, introduced in Angrist et al. (2021), offers a solution to the undersubscription problem.[41] IV VAM sidesteps underidentification of the β_j's with a model relating value-added to a lower-dimensional set of school characteristics, estimated by IV. This approach starts with a hypothetical school-level projection of value-added β_j on a $K \times 1$ vector of school characteristics M_j:

$$\beta_j = M_j'\varphi + \nu_j. \tag{40}$$

Equation (40) is a between-school model of the sort often used in hierarchical linear models for school effects (Raudenbush and Bryk, 1986). The characteristics M_j might include an OLS value-added coefficient α_j as well as other school attributes like sector or demographics. Coefficient vector φ captures the systematic relationship between these predictors and causal school quality. Residual ν_j, defined so that $E[M_j\nu_j] = 0$, reflects variation in school quality not explained by M_j. Hyperparameter $\sigma_\nu^2 \equiv Var(\nu_j)$ summarizes the extent of such residual variation.

Plugging the school-level projection (40) into the student-level causal model (22) yields:

$$Y_i = \tau_0 + M_{d(i)}'\varphi + \varepsilon_i + \nu_{d(i)}, \tag{41}$$

where $M_{d(i)} = \sum_j M_j D_{ij}$ is the vector of characteristics for student i's enrolled school. The first step of IV

[41] The IV VAM method simplifies and generalizes a parametric approach to hybrid estimation with undersubscription explored in Angrist et al. (2017).

VAM is to estimate equation (41) by 2SLS, instrumenting $M_{d(i)}$ with a vector of risk-adjusted offers for L oversubscribed schools, $\tilde{Z}_i = (Z_{i1} - p_{i1}, ..., Z_{iL} - p_{iL})$.

This 2SLS procedure generalizes the testing approach introduced in Section 5.3. When $M_{d(i)}$ consists only of an OLS school value-added estimate $\hat{\alpha}_{d(i)}$, the 2SLS estimate $\hat{\varphi}$ checks for forecast bias in the underlying OLS model, while the accompanying overidentification test assesses variation in the predictive value of OLS across lotteries. In more general cases, the 2SLS forecast coefficient describes the relationship between school characteristics M_j and value-added, and the overidentification test statistic can be used to construct an estimate of σ_ν^2 (which should be zero if M_j includes an unbiased OLS coefficient). The mechanics of the IV VAM estimation procedure and its connection to VAM specification tests are detailed further in Angrist et al. (2021).[42]

The second step of IV VAM constructs minimum MSE predictions of individual school quality given all available information, including the school characteristics M_j and estimates from the available lottery quasi-experiments. Let $\hat{\rho}$ denote the $L \times 1$ vector of coefficients from a regression of Y_i on the risk-adjusted offer vector $\tilde{Z}_i$, and let V_ρ denote the sampling covariance matrix of $\hat{\rho}$. The $J \times 1$ vector of IV VAM posteriors is given by:

$$\beta^* = \Pi'(\Pi\Pi' + V_\rho/\sigma_\nu^2)^{-1}\hat{\rho} + \left[I_J - \Pi'(\Pi\Pi' + V_\rho/\sigma_\nu^2)^{-1}\Pi\right] M\varphi. \tag{42}$$

Here M is a $J \times K$ matrix collecting characteristics M_j for all schools, I_J is the $J \times J$ identity matrix, and Π is an $L \times J$ matrix of first-stage coefficients from regressions of each of the J school attendance indicators, D_{ij}, on $\tilde{Z}_i$. In a special case with no undersubscription ($L = J$) and M_j equal to an OLS value-added coefficient, this formula collapses to the bivariate shrinkage formula (39).[43] With undersubscription, IV VAM posteriors combine a value-added forecast $M_j'\varphi$ for each school with reduced-form offer effects $\hat{\rho}$, accounting for offer compliance via the first-stage matrix Π. An EB implementation plugs OLS estimates of Π and V_ρ into (42) along with 2SLS estimates of φ and σ_ν^2, obtained from the first step of IV VAM.

5.5 Risk-Controlled Value-added Models (RC VAM)

An alternative strategy for using centralized assignment variation to estimate school value-added is the risk-controlled value-added model (RC VAM) of Angrist et al. (2021). This approach starts with the observation that assignment systems like the DA mechanism in Section 4 generate rich data on student preferences and priorities which may account for much of the non-random sorting across schools. RC VAM uses this data to bolster the selection-on-observables assumption in conventional VAM estimation, by adding functions of the assignment propensity scores p_{ij} to the control vector X_i in regression (32). In other words, instead of using centralized assignment information to generate instruments for school assignment, RC VAM uses this

[42] In cases where $Var(\nu_j) \neq 0$, the exclusion restriction underlying 2SLS estimation of (41) requires school offers to be uncorrelated with residual school quality $\nu_{d(i)}$, which is not guaranteed by independence of offers and potential outcomes. As detailed in Angrist et al. (2021), this makes IV VAM a special case of the “many invalid instruments” framework of Kolesár et al. (2015).

[43] See the Appendix to Angrist et al. (2021) for the details of this equivalence.

information to construct new control variables that help to mitigate selection bias in observational models.[44]

The selection-on-observables assumption underlying an RC VAM model with risk controls echoes that invoked in studies of college quality by Dale and Krueger (2002, 2014) and Mountjoy and Hickman (2020). These studies control for a student's college application portfolio and admissions offers to estimate the returns to enrollment at particular colleges, assuming that offer takeup decisions are as good as random. Similarly, RC VAM requires non-compliance with centralized assignment admission offers to be as good as random conditional on assignment risk and other observables. This connection is formalized in the following result of Angrist et al. (2021):

$$\varepsilon_i \perp\!\!\!\perp D_i | (p_i, X_i, Z_i) \implies \varepsilon_i \perp\!\!\!\perp D_i | (p_i, X_i), \tag{43}$$

where $D_i \in \{1, ..., J\}$ is the school attended by student i, $Z_i \in \{1, ..., J\}$ is i's school assignment, $p_i = (p_{i1}, ..., p_{iJ})$ is the vector of propensity scores for all schools, and ε_i is the student ability term from model (22). This result shows that if school enrollment is independent of ability among students with the same assignment risk, covariates, and offers, then enrollment is also independent of ability conditional on just risk and covariates—since offers are random conditional on risk, it is not necessary to control for offers once we've conditioned on the assignment propensity score. Knowledge of the propensity score therefore allows us to use the conditional randomness of offers to test the RC VAM selection-on-observables assumption (the right-hand side of (43)) rather than controlling directly for admission offers, even while motivating the RC VAM strategy by a Dale and Krueger-style assumption of random non-compliance with offers (the left-hand side of (43)).

Tests of the RC VAM identifying assumption for NYC middle and high schools appear in Figure 4. The results show that RC VAM estimates of NYC middle and high school quality are virtually unbiased, using the sample of Angrist et al. (2021) and same visual IV testing procedure as in Figure 3.[45] The first column of both panels shows again that uncontrolled VAMs, which effectively compare achievement levels, badly fail to predict the reduced-form effects of centralized school assignment. Remarkably, the second column shows that most of this selection bias is removed by adding assignment risk controls; when conventional VAM controls (demographics and lagged test scores) are further added the forecast coefficient is indistinguishable from one with all points tightly clustered around the 45-degree line. The omnibus test p-value for the middle and high school RC VAMs are 0.21 and 0.84, respectively. The seeming lack of omitted variables bias is especially impressive for high schools, where the NYC outcome (SAT scores) comes from a different test than the lagged score controls and thus may be more prone to omitted variables bias (a point made in a different context by Chetty et al. (2014b)). These results suggest that using information on assignment risk from centralized assignment systems is a promising strategy for mitigating selection bias in school VAMs.

[44] Abdulkadiroğlu et al. (2020) develop a related control-function approach that controls for preferences derived from random utility discrete-choice models fit to students' rank-ordered preference lists.

[45] Since there are many admissions instruments in NYC, Angrist et al. (2021) group them into 20 bins on the basis of the school's observational VAM estimate. Angrist et al. (2021) also find RC VAM to be virtually unbiased in a sample of Denver middle schools.

6 Conclusion: What Next for School Quality Measurement?

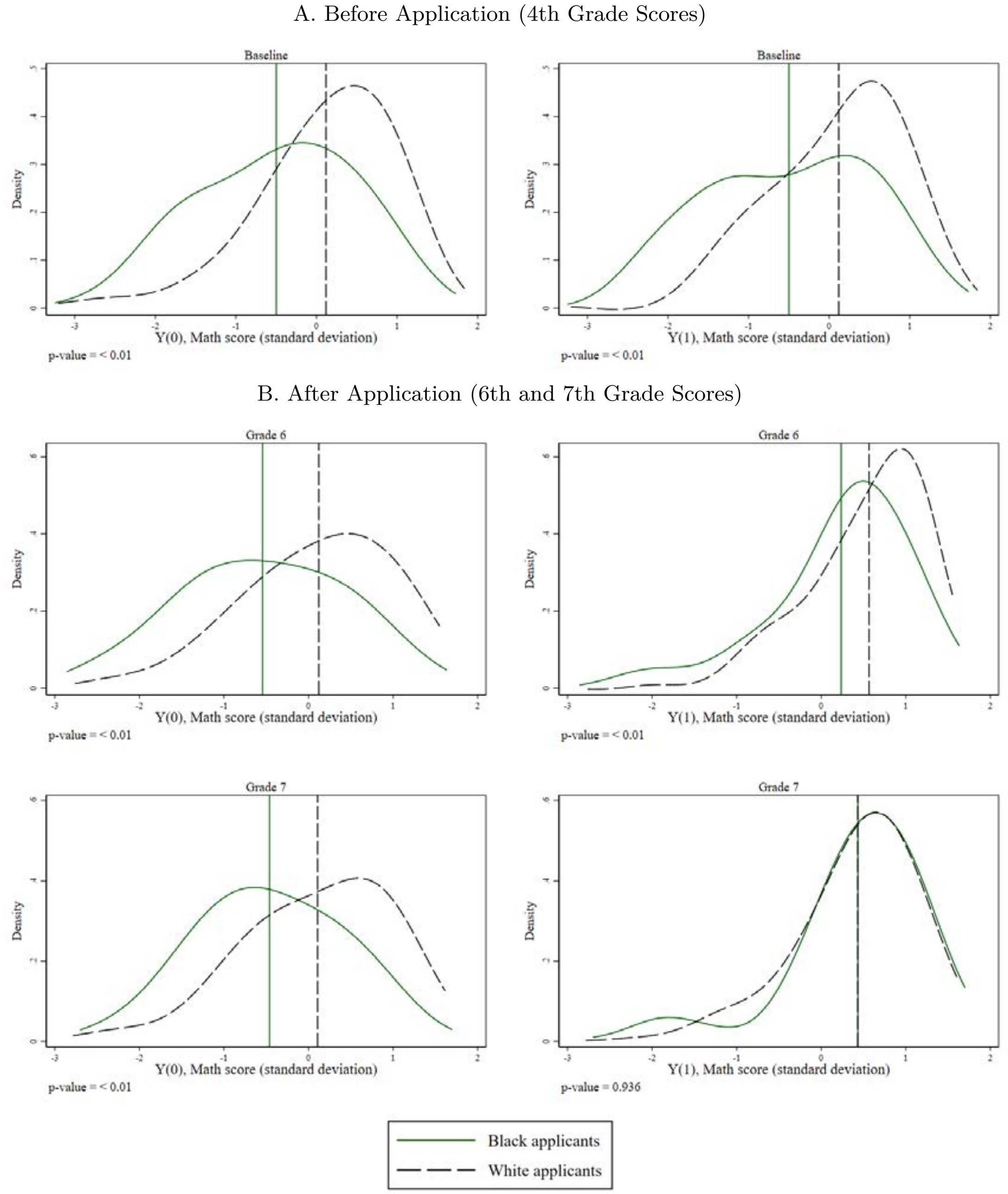
The increasing use of centralized assignment systems in American school districts offers new opportunities to apply the methods outlined in this chapter. Policymakers' and parents' growing demand for reliable measures of school effectiveness will likewise fuel such analysis. We conclude with a brief look at new directions this work might take.

Most of the research reviewed here focuses on achievement-based measures of school quality. Recent years have seen growing interest in school effects on outcomes other than student achievement. Early efforts in this direction include explorations of school effects on non-cognitive outcomes like absences, suspensions, and socioemotional development as well as longer-run outcomes like educational attainment, crime, political participation, employment, and earnings (Deming, 2011; Deming et al., 2014; Angrist et al., 2016a; Dobbie and Fryer, 2015; Abdulkadiroğlu et al., 2020; Dobbie and Fryer, 2020; Jackson et al., 2020; Beuermann et al., 2021; Cohodes and Feigenbaum, 2021). The links between achievement-based and alternative measures of value-added constitutes an important area for future research. Longer-term outcomes also raise new econometric issues, since the lagged-controls strategy used for conventional achievement-based VAMs is unavailable for something like earnings. Lottery-based methods may therefore be especially important when looking at longer-term effects.

The value-added models discussed in Section 5 posit a single causal effect of each school common to all students. In practice, school value-added may be heterogeneous – for example, a school's effect might be different for those attending in different years, or for students with different levels of preparation. And urban charter schools seem especially beneficial for students with low baseline test scores (Angrist et al., 2012; Frandsen and Lefgren, 2021; Chabrier et al., 2016). Match effects of this sort create new methodological challenges. In particular, recent work by Goldsmith-Pinkham et al. (2022) highlights problems with regression-based approaches in settings with many treatments and heterogeneous effects; match effects also complicate the interpretation of lottery-based tests for VAM validity, since these are sensitive to differences between LATEs and other treatment effect parameters in addition to selection bias. This suggest a role for more flexible, possibly non-parametric models beyond the linear 2SLS workhorse.

A third research frontier combines the school quality measurement tools discussed here with preference data generated by centralized assignment in an effort to understand the interplay between school choice and school effectiveness. Evidence on whether households choose schools based on causal impacts on student outcomes is mixed: (Rothstein, 2006; Abdulkadiroğlu et al., 2020) argue that school choice depends more on peer achievement than on causal school effects. Other work, however, suggests logistical barriers and seemingly complex or opaque choice systems choice systems prevent some households from selecting the most effective schools for their children (Walters, 2018; Bergman et al., 2020; Kapor et al., 2020). A better understanding of these issues should have the double payoff of improving both school accountability measures and educational outcomes.

Exhibits

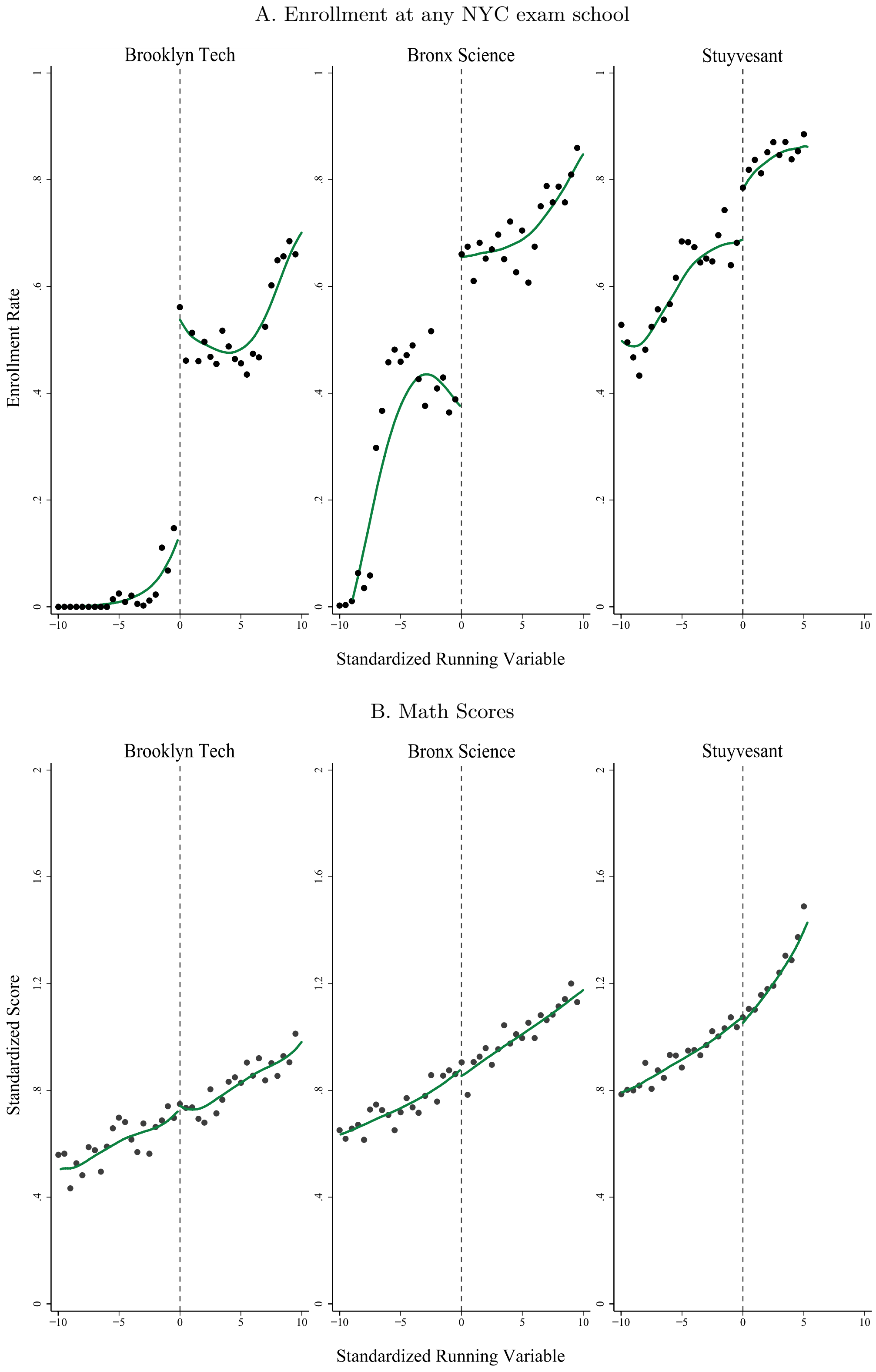


A. Before Application (4th Grade Scores)

B. After Application (6th and 7th Grade Scores)

Figure 1: Complier Distributions for Applicants to Massachusetts Urban Charters

Notes: This figure plots estimated distributions of untreated ($Y_i(0)$) and treated ($Y_i(1)$) potential outcomes for Black and white lottery compliers in the urban charter applicant sample. Distributions are estimated as described in Section 3.2, using the rule-of-thumb bandwidth described in footnote 10. Vertical lines show mean potential outcomes separately by race. p-values for testing the equality of white and Black distributions in each panel come from a weighted bootstrap procedure using the maximum absolute Black-white distance in estimated complier CDFs as the test statistic. CDFs in each bootstrap iteration are computed by estimating equation (11) by 2SLS with $g(X_i, Y_i) = 1\{Y_i \leq y\}$ for a grid of points y, weighting observations with *iid* exponential weights. The sample used here comes from Angrist et al. (2013).



A. Enrollment at any NYC exam school

B. Math Scores

Figure 2: First Stage and Reduced Form for NYC Exam School Admissions RD

Notes: Panel A shows NYC exam school enrollment rates for applicants to three exam schools, Brooklyn Tech, Bronx Science, and Stuyvesant, as a function of the distance of a student's admission score to each school's admission cutoff. Panel B shows corresponding reduced form impacts on Regents math standardized test scores. This figure shows the admission RD approach as applied in Abdulkadiroğlu et al. (2014).

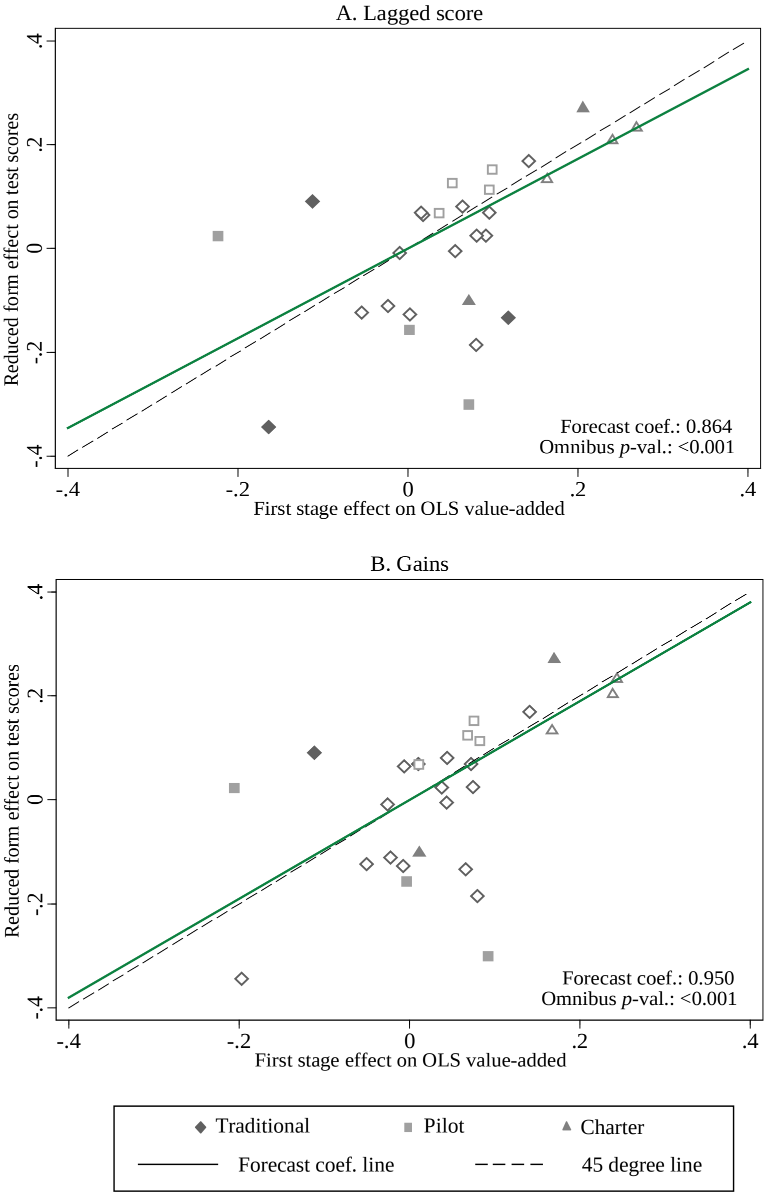


Figure 3: Visual IV Tests for VAM Bias

Notes: This figure plots lottery reduced-form estimates against value-added first stages from 28 middle school admission lotteries. Outcomes are standardized 6th grade math test scores. Schools are categorized as belonging to the charter, pilot, and traditional public sector. Filled markers indicate reduced form and first stage estimates that are significantly different from each other at the 10% level. The solid lines have slopes equal to the forecast coefficients in Table 7, while dashed lines indicate the 45-degree line. The sample used here comes from Angrist et al. (2017).

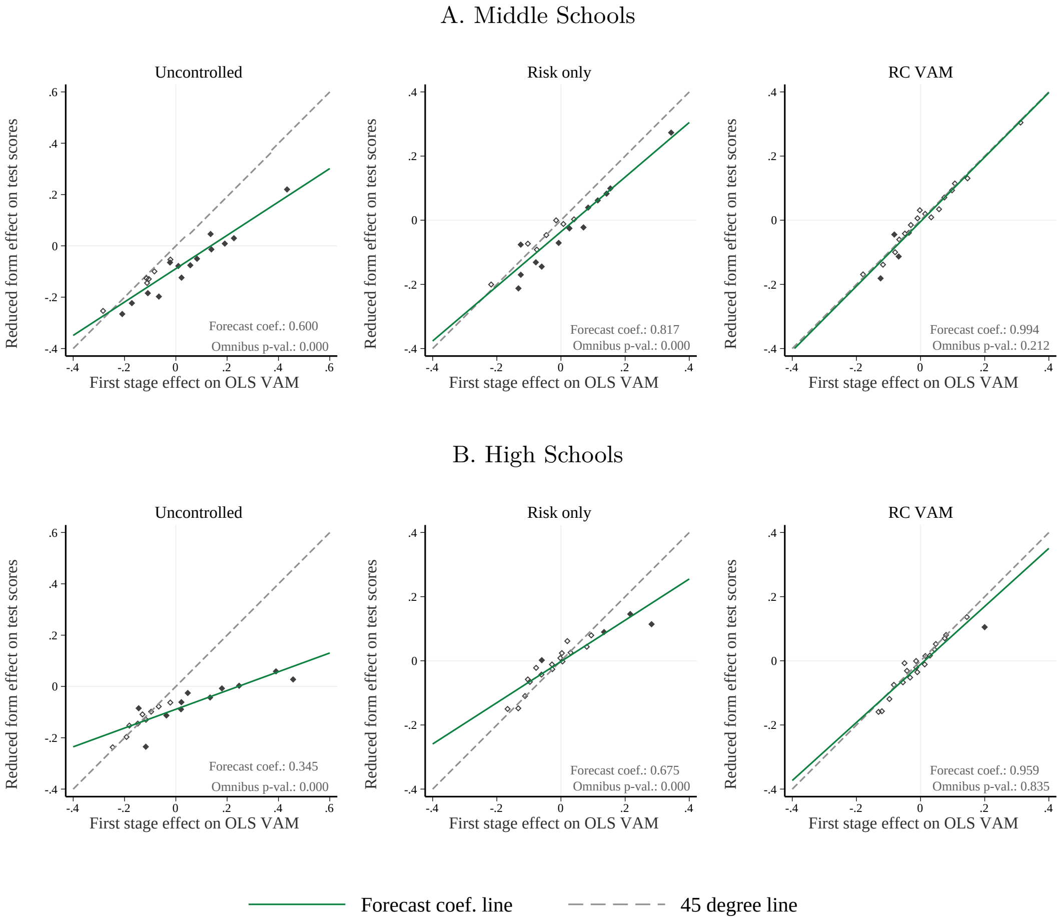


Figure 4: Visual IV Tests for VAM Bias

Notes: This figure plots reduced-form estimates against value-added first stages from each of 20 bins of school assignment indicators for NYC middle and high school samples. Outcomes are 6th grade math New York State Assessment scores for middle schools, and SAT math scores for high schools. Assignments are binned by ventile of the estimated conventional VAM. Filled markers indicate reduced form and first stage estimates that are significantly different from each other at the 10% level. The solid lines have slopes equal to the forecast coefficients in Table 2 of Angrist et al. (2021), while dashed lines indicate the 45-degree line. The sample used here comes from Angrist et al. (2021).

Table 1: Balance and Attrition for Massachusetts Urban Charter Lotteries

	Means		
	MA Urban (1)	Urban Applicant (2)	Balance Coefficient (3)
A. Balance			
Female	0.484	0.498	0.002 (0.017)
Black	0.201	0.479	-0.011 (0.015)
Hispanic	0.321	0.244	0.026 (0.014)
Asian	0.072	0.017	0.001 (0.005)
White	0.375	0.204	-0.007 (0.012)
Special education	0.200	0.176	-0.005 (0.013)
English language learner	0.161	0.103	0.003 (0.010)
Subsidized lunch status	0.688	0.688	0.008 (0.015)
Baseline math score	-0.416	-0.336	-0.020 (0.033)
Baseline English score	-0.454	-0.359	0.002 (0.035)
Joint p-value			0.694
B. Attrition			
Has outcome score	0.702	0.801	0.012 (0.010)
Observations	234,793	6,038	6,038

Notes: Columns (1) and (2) of this table report means of baseline characteristics and an indicator for having an outcome test score for students in Massachusetts urban school districts and urban charter lottery applicants, respectively. Column (3) reports coefficients from regressions of covariates on charter offer dummies, controlling for lottery risk set indicators. Robust standard errors are reported in parentheses. The joint p-value is from the test of the hypothesis that lottery offers are balanced on all baseline characteristics. The sample used here comes from Angrist et al. (2013).

Table 2: 2SLS Estimates for Massachusetts Urban Charter Schools

	Treatment Variable					
	Attendance Indicator				Years Attended	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)
Math score effects	0.329 (0.020)	0.454 (0.039)	0.407 (0.019)	0.579 (0.038)	0.236 (0.009)	0.314 (0.020)
First stage	0.567 (0.015)		0.551 (0.015)		1.017 (0.030)	
Non-charter outcome mean	-0.320		-0.268		-0.268	
Number of Applicants	4,281		4,590		4,590	
Sample Coverage	Application Year		All Years		All Years	
Sample Size	4,281		11,458		11,458	

Notes: This table reports OLS and 2SLS estimates of the effect of Boston charter middle school attendance on math test scores for applicants to urban charter schools in Massachusetts, as well as first stage estimates. Columns (1)-(2) define treatment as an indicator for enrolling in a charter school in the academic year following application and restricts the sample to test score outcomes from this application year. Columns (3)-(4) use the same treatment definition but pools post-lottery test scores for grades 4 through 7. Columns (5)-(6) use the full sample but define treatment as the number of years spent in a charter school by the outcome grade. All models control for lottery risk set indicators and student gender, race, special education, English language learner, subsidized lunch status, and grade and year indicators. Columns (1) and (2) report robust standard errors in parentheses. Standard errors are clustered by student in columns (3)-(6). The sample used here comes from Angrist et al. (2013).

Table 3: Characteristics of Lottery Compliers at Massachusetts Urban Charter Schools

	Compliers				
	Untreated (1)	Treated (2)	Pooled (3)	Always-takers (4)	Never-takers (5)
Female	0.506	0.510	0.508	0.539	0.463
	(0.023)	(0.021)	(0.016)	(0.024)	(0.017)
Black	0.401	0.380	0.390	0.623	0.490
	(0.022)	(0.021)	(0.016)	(0.023)	(0.017)
Hispanic	0.250	0.300	0.275	0.183	0.228
	(0.02)	(0.018)	(0.013)	(0.019)	(0.014)
Asian	0.022	0.024	0.023	0.004	0.024
	(0.007)	(0.005)	(0.004)	(0.003)	(0.005)
White	0.229	0.216	0.223	0.154	0.215
	(0.018)	(0.016)	(0.012)	(0.016)	(0.014)
Special education	0.190	0.181	0.186	0.158	0.177
	(0.018)	(0.016)	(0.012)	(0.018)	(0.013)
English language learner	0.143	0.148	0.145	0.054	0.088
	(0.015)	(0.013)	(0.010)	(0.011)	(0.010)
Subsidized lunch	0.689	0.705	0.697	0.698	0.666
	(0.021)	(0.019)	(0.014)	(0.022)	(0.016)
Baseline math score	-0.274	-0.312	-0.293	-0.394	-0.301
	(0.047)	(0.041)	(0.032)	(0.045)	(0.036)
Baseline English score	-0.352	-0.349	-0.350	-0.362	-0.299
	(0.050)	(0.043)	(0.033)	(0.046)	(0.038)
Share of sample			0.546	0.197	0.257

Notes: This table reports estimates of average baseline characteristics of compliers, always-takers, and never-takers among lottery applicants to urban charter schools in Massachusetts. Means are computed from 2SLS and OLS regressions that control for lottery risk set indicators, as described in Section 3.2. Robust standard errors are reported in parentheses. The sample used here comes from Angrist et al. (2013).

Table 4: Counterfactual School Destinies for Boston Charter Compliers

	Target sector					
	Proven providers		Expansion charters		Other charters	
	Z = 0	Z = 1	Z = 0	Z = 1	Z = 0	Z = 1
Destiny	(1)	(2)	(3)	(4)	(5)	(6)
Proven providers		1.000	-0.052 (0.038)		0.000 (0.024)	
Expansion charters	0.269 (0.046)			1.000	0.231 (0.034)	
Other charters	0.008 (0.026)		0.047 (0.026)			1.000
Traditional publics	0.528 (0.058)		0.694 (0.058)		0.529 (0.042)	
Pilots	0.180 (0.041)		0.174 (0.041)		0.118 (0.023)	

Notes: This table reports the share of untreated (Z=0) and treated (Z=1) compliers enrolled at particular fallback school types among applicants to Boston charter school lotteries. Destinies labelled at left are sectors enrolling treated and untreated compliers. Robust standard errors are reported in parentheses. The lottery sample used here comes from Cohodes et al. (2021).

Table 5: Multi-sector 2SLS Estimates for Boston Charter Schools

	Before charter expansion			After charter expansion			
		Estimates			Estimates		
	Non-charter mean (1)	Proven providers (2)	Other charters (3)	Non-charter mean (4)	Proven providers (5)	Expansion charters (6)	Other charters (7)
Math Score	0.117	0.320 (0.037)	0.183 (0.026)	-0.074	0.365 (0.070)	0.326 (0.074)	0.193 (0.055)
First stage							
Immediate offer		1.304 (0.067)	1.554 (0.047)		0.795 (0.054)	0.659 (0.046)	0.930 (0.052)
Waitlist offer		1.027 (0.050)	0.984 (0.061)		0.400 (0.048)	0.348 (0.041)	0.853 (0.071)

Notes: This table reports first stage effects of charter lottery offers on years of enrollment in charter schools and 2SLS estimates of the effects of charter school attendance on math test scores for multiple types of Boston charter middle schools. The sample stacks post-lottery test scores in grades five through eight. The endogenous variables are counts of years spent in the different charter types (pre-expansion proven providers, pre-expansion other charters, post-expansion proven providers, expansion schools, and post-expansion other charters). The instruments are immediate and waitlist lottery offer dummies for each school type. Immediate offer equals one for applicants offered seats on the day of the lottery. Waitlist offer equals one for applicants offered seats from the waitlist. Controls include lottery risk sets, as well as gender, race, ethnicity, a female-minority interaction, special education, English language learner, subsidized lunch status, and grade and year indicators. Standard errors, clustered by student, are reported in parentheses. The sample used here comes from Cohodes et al. (2021).

Table 6: Alternative IV Strategies for Denver Charter Effects

	Instruments		
	Offer	First Choice	Qualification
	(1)	(2)	(3)
Math score	0.417 (0.050)	0.515 (0.064)	0.379 (0.092)
First stage	0.443 (0.024)	0.347 (0.022)	0.457 (0.021)
Risk controls	DA Score	first-choice risk sets	preference risk sets
Equivalent sample increase vs. column (1)		1.64	3.46
Observations	2,099	2,222	3,502

Notes: This table reports IV estimates of the effects of charter school attendance for students in Denver using strategies based on centralized school assignment. The first row compares alternative 2SLS estimates of charter attendance effects on math scores. The second row reports the corresponding first stage estimates. Column (1) instruments charter attendance with centralized assignment to a charter school, controlling for percentiles of the simulated charter assignment propensity score and other baseline covariates. Column (2) instruments charter attendance with a charter first-choice assignment instrument, controlling for first-choice fixed effects and other baseline covariates. Column (3) instruments charter attendance with a charter qualification instrument, controlling for preference fixed effects and other baseline covariates. The fourth row reports the sample size increase needed to achieve a precision gain equivalent to the gain from using the any-charter offer instrument. Robust standard errors are reported in parentheses. The sample used here is as in Abdulkadiroğlu et al. (2017).

Table 7: VAM Bias Tests for Boston Schools

	Value-added model		
	Uncontrolled (1)	Lagged score (2)	Gains (3)
Forecast coefficient	0.396 (0.056)	0.864 (0.075)	0.950 (0.084)
p-values:			
Forecast bias	<0.001	0.071	0.554
Overidentification	<0.001	0.003	0.006
Omnibus	<0.001	<0.001	<0.001

Notes: This table reports the results of tests for bias in conventional VAMs using sixth-grade math scores for students in Boston. The uncontrolled model includes only year-of-test indicators as controls. The lagged score VAM includes cubic polynomials in baseline math and ELA scores, along with indicators for application year, sex, race, subsidized lunch, special education, English language learner status, and counts of baseline absences and suspensions. The gains VAM drops the lagged score controls and uses score growth from baseline as the outcome. Forecast coefficients are from IV regressions of test scores on fitted values from conventional VAMs, instrumenting fitted values with lottery offer indicators. The IV models are estimated via an asymptotically efficient GMM procedure and control for assignment strata fixed effects, demographic variables, and lagged scores. The forecast bias test checks whether the forecast coefficient equals 1, and the overidentificiation test checks the IV model's overidentifying restrictions. The omnibus test combines forecast bias and overidentifying restrictions. Standard errors are reported in parentheses. The sample used here comes from Angrist et al. (2017).

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292.

——— (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Abaluck, J., M. M. C. Bravo, P. Hull, and A. Starc (2021): "Mortality Effects and Choice Across Private Health Insurance Plans," *Quarterly Journal of Economics*, 136, 1557–1610.

Abdulkadiroğlu, A., J. Angrist, and P. Pathak (2014): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82, 137–196.

Abdulkadiroğlu, A. and T. Andersson (2022): "School Choice," NBER Working Paper 29822.

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

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

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

——— (2021): "Breaking Ties: Regression Discontinuity Design Meets Market Design," *Forthcoming, Econometrica.*

Abdulkadiroğlu, A., W. Hu, and P. Pathak (2013): "Small High Schools and Student Achievement: Lottery-Based Evidence from New York City," NBER Working Paper 19576.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020): "Do Parents Value School Effectiveness?" *American Economic Review*, 110, 1502–39.

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

Allcott, H. (2015): "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 130, 1117–1165.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Anelli, M. (2020): "The Returns to Elite University Education: a Quasi-Experimental Analysis," *Journal of the European Economic Association*, 18, 2824–2868.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92, 1535–1558.

Angrist, J., P. Hull, P. Pathak, and C. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 132, 871–919.

Angrist, J. and M. Kolesár (2021): "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV," Working Paper 29417, National Bureau of Economic Research.

Angrist, J. D. (2004): "Treatment Effect Heterogeneity in Theory and Practice," *Economic Journal*, 114, C52–C83.

Angrist, J. D., S. R. Cohodes, S. M. Dynarski, P. A. Pathak, and C. R. Walters (2016a): "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice," *Journal of Labor Economics*, 34, 275–318.

Angrist, J. D., S. M. Dynarski, T. J. Kane, P. A. Pathak, and C. R. Walters (2010): "Inputs and Impacts in Charter Schools: KIPP Lynn," *American Economic Review: Papers & Proceedings*, 100, 239–243.

——— (2012): "Who Benefits from KIPP?" *Journal of policy Analysis and Management*, 31, 837–860.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2016b): "Interpreting Tests of School VAM Validity," *American Economic Review: Papers & Proceedings*, 106, 388–392.

——— (2021): "Credible School Value-Added with Undersubscribed School Lotteries," MIT Blueprint Labs Working Paper. Forthcoming, *Review of Economics and Statistics*.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

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

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2020): "Robust Empirical Bayes Confidence Intervals," *arXiv preprint arXiv:2004.03448*.

Baude, P. L., M. Casey, E. A. Hanushek, G. R. Phelan, and S. G. Rivkin (2020): "The Evolution of Charter School Quality," *Economica*, 87, 158–189.

Behaghel, L., B. Crépon, and M. Gurgand (2013): "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial," Discussion Paper 7447, Institute for the Study of Labor (IZA), Bonn, Germany.

Behaghel, L., C. de Chaisemartin, and M. Gurgand (2017): "Ready for Boarding? The Effects of a Boarding School for Disadvantaged Students," *American Economic Journal: Applied Economics*, 9, 140–64.

Bergman, P., E. W. Chan, and A. Kapor (2020): "Housing Search Frictions: Evidence from Detailed Search Data and a Field Experiment," NBER Working Paper 27209.

Bergman, P. and M. J. Hill (2018): "The Effects of Making Performance Information Public: Regression Discontinuity Evidence from Los Angeles Teachers," *Economics of Education Review*, 66, 104–113.

Beuermann, D., C. K. Jackson, L. Navarro-Sola, and F. Pardo (2021): "What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output," NBER Working Paper 25342.

Beuermann, D. W. and C. K. Jackson (2022): "The Short- and Long-Run Effects of Attending the Schools that Parents Prefer," *Journal of Human Resources*, 57, 725–746.

Bhuller, M. and H. Sigstad (2022): "2SLS with Multiple Treatments," *arXiv preprint arXiv:2205.07836*.

Bleemer, Z. and A. Mehta (2022): "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major," *American Economic Journal: Applied Economics*, 14, 1–22.

Bloom, H. S. (1984): "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 8, 225–246.

Bloom, H. S. and R. Unterman (2014): "Can Small High Schools of Choice Improve Educational Prospects for Disadvantaged Students?" *Journal of Policy Analysis and Management*, 33, 290–319.

Borusyak, K. and P. Hull (2020): "Non-Random Exposure to Exogenous Shocks: Theory and Applications," NBER Working Paper 27845.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039.

Brown, L. D. (2008): "In-Season Prediction of Batting Averages: A Field Test of Empirical Bayes and Bayes Methodologies," *The Annals of Applied Statistics*, 2, 113–152.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

Campos, C. Q. and C. Kearns (2022): "The Impacts of Neighborhood School Choice: Evidence from Los Angeles' Zones of Choice," Mimeo, University of Chicago.

Card, D. and L. Giuliano (2016): "Can Tracking Raise the Test Scores of High-Ability Minority Students?" *American Economic Review*, 106, 2783–2816.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2016): "Inference in Regression Discontinuity Designs under Local Randomization," *The Stata Journal*, 16, 331–367.

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

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2563.

——— (2014b): "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R. and N. Hendren (2018): "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," *Quarterly Journal of Economics*, 133, 1107–1162.

Chingos, M. M. and P. E. Peterson (2015): "Experimentally Estimated Impacts of School Vouchers on College Enrollment and Degree Attainment," *Journal of Public Economics*, 122, 1–12.

Clark, M. A., P. M. Gleason, C. C. Tuttle, and M. K. Silverberg (2015): "Do Charter Schools Improve Student Achievement?" *Educational Evaluation and Policy Analysis*, 37, 419–436.

Cohodes, S. and J. J. Feigenbaum (2021): "Why Does Education Increase Voting? Evidence from Bostonâs Charter Schools," NBER Working Paper 29308.

Cohodes, S. R., E. M. Setren, and C. R. Walters (2021): "Can Successful Schools Replicate? Scaling up Boston's Charter School Sector," *American Economic Journal: Economic Policy*, 13, 138–67.

Coleman, J. S. (1966): *Equality of Educational Opportunity*, Washington, DC: Government Printing Office.

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

Curto, V. E. and R. G. Fryer (2014): "The Potential of Urban Boarding Schools for the Poor: Evidence from SEED," *Journal of Labor Economics*, 32, 65–93.

Dale, S. B. and A. B. Krueger (2002): "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," *Quarterly Journal of Economics*, 117, 1491–1527.

——— (2014): "Estimating the Effects of College Characteristics over the Career Using Administrative Earnings Data," *Journal of Human Resources*, 49, 323–358.

Davis, M. and B. Heller (2019): "No Excuses Charter Schools and College Enrollment: New Evidence from a High School Network in Chicago," *Education Finance and Policy*, 14, 414–440.

de Chaisemartin, C. and L. Behaghel (2020): "Estimating the Effect of Treatments Allocated by Randomized Waiting Lists," *Econometrica*, 88, 1453–1477.

de Roux, N. and E. Riehl (2022): "Do College Students Benefit from Placement into Higher-Achieving Classes?" *Journal of Public Economics*, 210, 104669.

Deming, D. (2014): "Using School Choice Lotteries to Test Measures of School Effectiveness," *American Economic Review: Papers & Proceedings*, 104, 406–411.

Deming, D. J. (2011): "Better Schools, Less Crime?" *Quarterly Journal of Economics*, 126, 2063–2115.

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

Dobbie, W. and R. Fryer (2014): "The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools," *American Economic Journal: Applied Economics*, 6, 58–75.

Dobbie, W. and R. G. Fryer (2011): "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, 3, 158–87.

——— (2013): "Getting Beneath the Veil of Effective Schools: Evidence from New York City," *American Economic Journal: Applied Economics*, 5, 28–60.

——— (2015): "The Medium-Term Impacts of High-Achieving Charter schools," *Journal of Political Economy*, 123, 985–1037.

——— (2020): "Charter Schools and Labor Market Outcomes," *Journal of Labor Economics*, 38, 915–957.

Draper, N. and H. Smith (1998): *Applied Regression Analysis*, Wiley, third ed.

Dustan, A., A. de Janvry, and E. Sadoulet (2017): "Flourish or Fail? The Risky Reward of Elite High School Admission in Mexico City," *Journal of Human Resources*, 52, 756–799.

Efron, B. (2012): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1, Cambridge University Press.

——— (2016): "Empirical Bayes Deconvolution Estimates," *Biometrika*, 103, 1–20.

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs," *Journal of Political Economy*, 123, 413–443.

Engberg, J., D. Epple, J. Imbrogno, H. Sieg, and R. Zimmer (2014): "Evaluating Education Programs That Have Lotteried Admission and Selective Attrition," *Journal of Labor Economics*, 32, 27–63.

Feller, A., T. Grindal, L. Miratrix, and L. C. Page (2016): "Compared to what? Variation in the impacts of early childhood education by alternative care type," *The Annals of Applied Statistics*, 10, 1245 – 1285.

Figlio, D. and C. M. D. Hart (2014): "Competitive Effects of Means-Tested School Vouchers," *American Economic Journal: Applied Economics*, 6, 133–56.

Frandsen, B. R. and L. J. Lefgren (2021): "Partial Identification of the Distribution of Treatment Effects with an Application to the Knowledge is Power Program (KIPP)," *Quantitative Economics*, 12, 143–171.

Gale, D. and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69, 9–15.

Gilraine, M., J. Gu, and R. McMillan (2020): "A New Method for Estimating Teacher Value-added," NBER Working Paper 27094.

Gilraine, M., U. Petronijevic, and J. D. Singleton (2021): "Horizontal Differentiation and the Policy Effect of Charter Schools," *American Economic Journal: Economic Policy*, 13, 239–76.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2022): "Contamination Bias in Linear Regressions," NBER Working Paper 30108.

Gu, J. and R. Koenker (2020): "Invidious Comparisons: Ranking and Selection as Compound Decisions," *arXiv preprint arXiv:2012.12550*.

Hasan, S. and A. Kumar (2019): "Digitization and Divergence: Online School Ratings and Segregation in America," SSRN Working Paper.

Hastings, J. S., T. J. Kane, and D. O. Staiger (2009): "Heterogeneous Preferences and the Efficacy of Public School Choice," Working Paper, Yale University.

Hastings, J. S., C. A. Neilson, and S. D. Zimmerman (2019): "Are Some Degrees Worth More than Others? Evidence from College Admission Cutoffs in Chile," NBER Working Paper 19241.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

Heinesen, E. (2018): "Admission to Higher Education Programmes and Student Educational Outcomes and Earnings–Evidence from Denmark," *Economics of Education Review*, 63, 1–19.

Hoekstra, M. (2009): "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach," *Review of Economics and Statistics*, 91, 717–724.

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

Hoxby, C. M., S. Murarka, and J. Kang (2009): "How New York City's Charter Schools Affect Achievement," Working Paper.

Huber, M. and G. Mellace (2015): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411.

Imbens, G. and K. Kalyanaraman (2011): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020): "School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment," *American Economic Review: Insights*, 2, 491–508.

James, W. and C. Stein (1961): "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.

Jia, R. and H. Li (2021): "Just Above the Exam Cutoff Score: Elite College Admission and Wages in China," *Journal of Public Economics*, 196, 104371.

Kane, T. J. and D. O. Staiger (2008): "Estimating Teacher Impacts on Student Achievement: an Experimental Evaluation," NBER Working Paper 14607.

Kapor, A. J., C. A. Neilson, and S. D. Zimmerman (2020): "Heterogeneous Beliefs and School Choice Mechanisms," *American Economic Review*, 110, 1274–1315.

Kiefer, J. and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887 – 906.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016): "Field of Study, Earnings, and Self-Selection," *The Quarterly Journal of Economics*, 131, 1057–1111.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-Out Estimation of Variance Components," *Econometrica*, 88, 1859–1898.

Kline, P. and C. R. Walters (2016): "Evaluating Public Programs with Close Substitutes: The Case of Head Start*," *The Quarterly Journal of Economics*, 131, 1795–1848.

——— (2019): "On Heckits, LATE, and Numerical Equivalence," *Econometrica*, 87, 677–696.

Kline, P. M., E. K. Rose, and C. R. Walters (2021): "Systemic Discrimination Among Large US Employers," NBER Working Paper 29053. Forthcoming, *Quarterly Journal of Economics.*

Kline, P. M. and C. R. Walters (2021): "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 89, 765–792.

Koenker, R. and I. Mizera (2014): "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules," *Journal of the American Statistical Association*, 109, 674–685.

Kolesár, M. (2013): "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity," Working paper.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015): "Identification and Inference With Many Invalid Instruments," *Journal of Business & Economic Statistics*, 33, 474–484.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lee, S., M. Niederle, and N. Kang (2014): "Do Single-Sex Schools Make Girls More Competitive?" *Economics Letters*, 124, 474–477.

Lee, Y. and N. Nakazawa (2021): "Does Single-Sex Schooling Help or Hurt Labor Market Outcomes? Evidence from a Natural Experiment in South Korea," Working paper.

List, J. A. (2021): *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, Penguin.

List, J. A., D. Suskind, and L. H. Supplee (2021): *The Scale-Up Effect in Early Childhood and Public Policy: Why Interventions Lose Impact at Scale and What we Can Do About It*, Routledge.

Lucas, A. and I. Mbiti (2014): "Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya," *American Economic Journal: Applied Economics*, 6, 234–263.

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

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, 111, 3663–98.

Morris, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55.

Mountjoy, J. and B. Hickman (2020): "The Returns to College(s): Estimating Value-Added and Match Effects in Higher Education," Becker Friedman Institute Working Paper.

Narita, Y. (2016): "(Non)Randomization: A Theory of Quasi-Experimental Evaluation of School Quality," Cowles Foundation Discussion Paper No. 2056.

Oosterbeek, H. and N. R. I. de Wolf (2021): "Heterogeneous Effects of Comprehensive vs. Single-Track Academic Schools: Evidence from Admission Lotteries," Working paper.

Pop-Eleches, C. and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103, 1289–1324.

Raudenbush, S. and A. S. Bryk (1986): "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 59, 1–17.

Robbins, H. (1956): "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 3.1, 157–163.

Rockoff, J. and L. J. Turner (2010): "Short-Run Impacts of Accountability on School Quality," *American Economic Journal: Economic Policy*, 2, 119–47.

Romero, M. and A. Singh (2021): "The Incidence of Affirmative Action: Evidence from Quotas in Private Schools in India," Working paper.

Rosenbaum, P. R. and D. B. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.

Roth, A. E. and M. A. O. Sotomayor (1990): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs.

Rothstein, J. (2010): “Teacher quality in Educational Production: Tracking, Decay, and Student Achievement,” *Quarterly Journal of Economics*, 125, 175–214.

Rothstein, J. M. (2006): “Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions,” *American Economic Review*, 96, 1333–1350.

Rothstein, R. (2004): *Class and Schools: Using Social, Economic, and Educational Reform to Close the BlackâWhite Achievement Gap*, New York: Teachers College Press.

Roy, A. D. (1951): “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135–146.

Sargan, J. (1958): “The Estimation of Economic Relationships using Instrumental Variables,” *Econometrica*, 26, 393–415.

Sekhri, S. (2020): “Prestige Matters: Wage Premium and Value Addition in Elite Colleges,” *American Economic Journal: Applied Economics*, 12, 207–25.

Setren, E. (2021): “Targeted vs. General Education Investments Evidence from Special Education and English Language Learners in Boston Charter Schools,” *Journal of Human Resources*, 56, 1073–1112.

Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Wald, A. (1940): “The Fitting of Straight Lines if Both Variables are Subject to Error,” *The Annals of Mathematical Statistics*, 11, 284–300.

Walters, C. R. (2018): “The Demand for Effective Charter Schools,” *Journal of Political Economy*, 126.

Zellner, A. and H. Theil (1962): “Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations,” *Econometrica*, 30, 54–78.

Zhang, H. (2014): “The Mirage of Elite Schools: Evidence from Lottery-based School Admissions in China,” Working Paper.

Zimmerman, S. D. (2014): “The Returns to College Admission for Academically Marginal Students,” *Journal of Labor Economics*, 32, 711–754.

——— (2019): “Elite Colleges and Upward Mobility to Top Jobs and Top Incomes,” *American Economic Review*, 109, 1–47.