Econometrica, Vol. 90, No. 6 (November, 2022), 2509–2539

EMPIRICAL STRATEGIES IN ECONOMICS: ILLUMINATING THE PATH FROM CAUSE TO EFFECT

JOSHUA D. ANGRIST
Department of Economics, MIT and NBER

The view that empirical strategies in economics should be transparent and credible now goes almost without saying. By revealing for whom particular instrumental variables (IV) estimates are valid, the local average treatment effects (LATE) framework helped make this so. This lecture uses empirical examples, mostly involving effects of charter and exam school attendance, to illustrate the value of the LATE framework for causal inference. LATE distinguishes independence conditions satisfied by random assignment from more controversial exclusion restrictions. A surprising exclusion restriction is shown to explain why enrollment at Chicago exam schools reduces student achievement. I also make two broader points: IV exclusion restrictions formalize commitment to clear and consistent explanations of reduced-form causal effects; the credibility revolution in applied econometrics owes at least as much to compelling empirical analyses as to methodological insights.

TO MEASURE THE EFFECT OF GOOD OR BAD WATER SUPPLY, it is requisite to find two classes of inhabitants living at the same level, moving in equal space, enjoying an equal share of the means of subsistence, engaged in the same pursuits, but differing in this respect—that one drinks water from Battersea, the other from Kew ... But of such *experimenta crucis* the circumstances of London do not admit.
—William Farr (1853, *Weekly Return of Births and Deaths in London*)

The experiment ... was on the grandest scale. No fewer than 300,000 people of both sexes, of every age and occupation, and of every rank and station, from gentle-folks down to the very poor, were divided into two groups without their choice, and, in some cases, without their knowledge; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water free from such impurity.
—John Snow (1855, *On the Mode of Communication of Cholera*, 2nd ed.)

1. INTRODUCTION

In a chapter in the *Handbook of Labor Economics*, Alan Krueger and I employed the phrase "empirical strategy" to describe econometric analysis of natural experiments like the one physician John Snow (1855) used to show that cholera is a waterborne illness. The Handbook volume in question (Ashenfelter and Card (1999)) was edited by two of my Princeton Ph.D. thesis advisors, Orley Ashenfelter and David Card, leaders of the effort to bring empirical strategies like Snow's into the econometric mainstream. Ashenfelter and Card's quest for an empirical strategy to capture the causal effects of government

Joshua D. Angrist: angrist@mit.edu

This is a revised version of my recorded Nobel Memorial Lecture (https://www.youtube.com/watch?v=DV4IjRfjTs0&t=1298s) posted December 8, 2021. Many thanks to Noa Benveniste, Jimmy Chin and Vendela Norman for their help preparing this lecture and to Noam Angrist, Hank Farber, Peter Ganong, Guido Imbens, Michal Kolesar, and Parag Pathak for comments. Thanks also go to my coauthors and Blueprint Labs (https://blueprintlabs.mit.edu/research/) colleagues, from whom I've learned more than I can say. Special thanks are due to my co-laureates, David Card and Guido Imbens, for their guidance and partnership. We three share a debt to our absent friend, Alan Krueger, with whom we collaborated so fruitfully. This lecture discusses empirical findings from joint work with Atila Abdulkadiroğlu, Sarah Cohodes, Sue Dynarski, Bill Evans, Iván Fernández-Val, Tom Kane, Victor Lavy, Yusuke Narita, Parag Pathak, Chris Walters, and Román Zárate.

https://doi.org/10.3982/ECTA20640

training programs inspired me and others at Princeton to explore the econometrics of program and policy evaluation.[1]

An empirical strategy for program or policy evaluation is a research plan that encompasses data collection, identification, and estimation. As Krueger and I used it, the term *identification* is shorthand for research design. The prize I share with David Card and Guido Imbens reflects the prominent role research design has come to play in modern economics. A randomized clinical trial (RCT) is the simplest and most powerful research design. Random assignment ensures that treatment and control groups are comparable in the absence of treatment, so post-treatment differences in average outcomes reflect only the treatment effect. Not surprisingly, though also not without resistance, RCTs have come to be both an aspiration and a benchmark for empirical strategies in economics.[2]

Krueger's (1999) class size study illustrates the power of RCTs to generate clear findings, and to set a standard by which non-randomized investigations are measured. The effects of reduced grade-school class size have long preoccupied economists interested in the education production function. For decades, Hanushek (1986, 1996) and others had argued that expenditure on education inputs, like smaller classes, is largely uncorrelated with higher achievement. Krueger's analysis of the Tennessee STAR class size RCT (which addressed econometric problems related to attrition, clustering, and noncompliance) shows strong, robust causal effects of class size reductions on learning. These findings have been integral to the school resources debate ever since. At the same time, RCTs on the scale of Tennessee STAR remain costly and time-consuming. Such ambitious social experiments are still unusual.

I see instrumental variables (IV) methods and regression discontinuity (RD) designs as the next best thing to an RCT when practical considerations inhibit use of experimental random assignment. In applications of IV and RD, causal variables of interest (like class size) are often referred to as treatment variables. This terminology evokes an analogy with RCTs that assign one or more treatments randomly. Like RCTs, compelling IV and RD strategies exploit an applied econometrician's understanding of the mechanisms that determine treatment assignment. It's this understanding that give IV and RD their causality-revealing power.

RD applications exploit assignment schemes in which treatment is allocated according to whether a classification variable (today called a *running variable*) clears a cutoff. As RD pioneers recognized, the method does not require the treatment variable whose causes we seek to switch fully on or off at the cutoff: fruitful RD requires only that the conditional mean of this variable jump at the cutoff.[3] Probabilistic manipulation leads to the use of

[1] Their quest began in Ashenfelter (1974, 1978) and Ashenfelter and Card (1985). A few years ahead of me, Ashenfelter student Robert J. LaLonde had shown how difficult the search was likely to be (LaLonde (1986)). Orley Ashenfelter not only brought me to Princeton and arranged to fund my studies (dayenu!), he suggested my thesis topic. Ashenfelter kicked off a Graduate Labor Economics class in 1986 by mentioning an intriguing study: Hearst, Newman, and Hulley (1986) compares the death rates of men with low and high draft lottery numbers as a gauge of the long-term health consequences of conscription. "Someone should do that for their earnings," quoth Orley. I went from class to the library, embarking on my first attempt to answer causal questions using observational data (https://www.nobelprize.org/uploads/2021/10/advanced-economicsciencesprize2021.pdf). Farr and Snow in the epigraph are quoted in Johnson (2006).

[2] The 2019 Economics Nobel awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer celebrates the rise of RCTs in development economics.

[3] RD originated in work by psychologists Donald Campbell and D.L. Thistlethwaite (Thistlethwaite and Campbell (1960)). Econometric RD pioneers Goldberger (1972) and Barnow (1972) discussed hypothetical applications of RD to evaluation of the then-nascent Head Start program. Cook (2008) and Lee and Lemieux (2010) sketch the intellectual history of RD.

discontinuities in the rate at which treatment is assigned as instrumental variables for treatment received. This sort of RD design is said to be *fuzzy*. But, as Steve Pischke and I wrote in our first book (Angrist and Pischke (2009)), "fuzzy RD is IV."

The first RD application to which I contributed is Angrist and Lavy (1999), a study that exploits the rule used in Israeli elementary schools to determine class size. In the 1990s, Israeli classes were large, with up to 40 students–the statutory maximum. In schools with 40 students in the same grade, students were likely to be seated in a class of 40. With another child added to the cohort, making 41, the cohort was likely to be split into two much smaller classes. This leads to the Maimonides' Rule research design, so named because the 12th Century philosopher and Torah scholar Maimonides proposed a maximum class size of 40.[4]

Figure 1 plots Israeli fourth grade class sizes as a function of contemporaneous fourth grade enrollment, overlaid with the class size prescribed by Maimonides Rule. The fit isn't perfect—it's this feature that makes our use of Maimonides' Rule a fuzzy RD design and necessitates use of IV. But the gist of the thing is a marked drop in average class size at

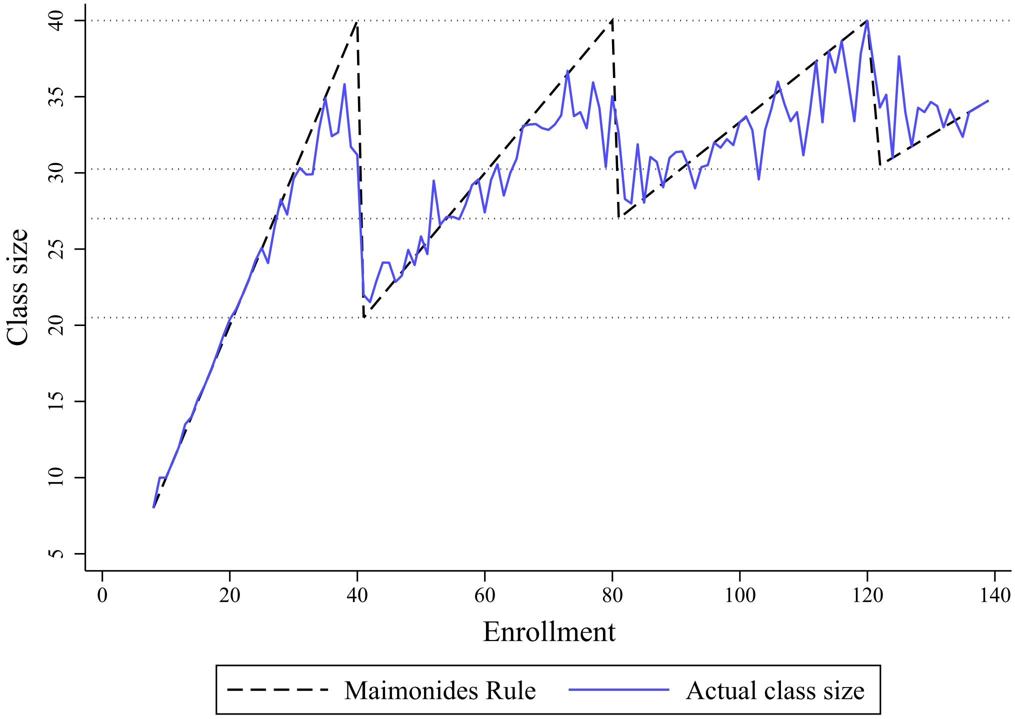


FIGURE 1.—Fourth grade class size by initial enrollment, actual average size and as predicted by Maimonides' Rule. *Notes*: Adapted from Angrist and Lavy (1999); data from Israel in 1991.

[4] The Rule is from Chapter II of "Laws Concerning the Study of Torah" in Book I of Maimonides' *Mishneh Torah*. Maimonides' proposal is founded on the Talmud, though the great sage appears to have taken liberties in favoring a cutoff of 40 over 50. The Talmud limits class size as follows: "The number of pupils assigned to each teacher is twenty-five. If there are fifty, we appoint two teachers. If there are forty, we appoint an assistant, at the expense of the town" (English translation on page 214 of Epstein (1976)).

multiples of 40 (the relevant cutoffs), as predicted by the Rule. As it turns out, these drops in class size are reflected in jumps in fourth (and fifth) grade test scores.[5]

Lavy and I implemented the Maimonides' Rule IV research design in a two-stage least squares (2SLS) set-up that can be described as follows. Writing f_j for predicted 4th grade class size at school j, Rule-based enrollment is:

$$f_j = \frac{r_j}{\left[\operatorname{int}\left((r_j - 1)/40\right) + 1\right]},$$

where r_j is the number of 4th graders at school j and $\operatorname{int}(x)$ is the largest integer less than or equal to x. The first-stage effect of instrumental variable f_j on class size is estimated by fitting:

$$s_{ij} = \pi f_j + h_1(r_j) + \delta_1' X_{ij} + \varepsilon_{ij}, \tag{1}$$

where s_{ij} is the class size experienced by student i enrolled in school j, X_{ij} is a vector of student and school characteristics, f_j and r_j are as defined above, and ε_{ij} is a regression error term. Second-stage models can be written:

$$y_{ij} = \beta s_{ij} + h_2(r_j) + \delta_2' X_{ij} + \eta_{ij}, \tag{2}$$

where β is the causal effect of interest and η_{ij} is the random part of potential achievement. Both first and second stage control for polynomial functions of the running variable, denoted by $h_1(r_j)$ in the first stage and $h_2(r_j)$ in the second. This IV model is identified because f_j is not only a nonlinear function of enrollment, it's discontinuous.

Angrist and Lavy (1999) uses the local average treatment effects (LATE) framework to interpret estimates based on (1) and (2) in a world of heterogeneous potential outcomes. Specifically, we showed that Rule-based IV estimates capture average causal effects for students pushed into smaller classes by Maimonides' Rule, for students attending classes that would be seen as unusually large by US standards. Yet, when converted to standard deviation units, the resulting class size effects are remarkably close to those reported by Krueger (1999). Following a suggestion from Caroline Hoxby, we also reported an analysis in samples limited to applicants close to Maimonides' Rule cutoffs, a simple nonparametric strategy.[6] Shortly thereafter, Hahn, Todd, and Van der Klaauw (2001) formalized the LATE interpretation of nonparametric fuzzy RD. Applications of this new approach to IV and RD, initially isolated, have since blossomed.

This lecture illustrates the power of IV and RD empirical strategies to uncover new causal knowledge through a series of examples. Most of these examples concern causal effects of attendance at schools of various kinds. Research on school effects highlights key features of the LATE framework, including an extension to causal effects on distributions. The extension shows how urban charter school attendance closes Black–White achievement gaps. Exclusion restrictions are typically the most controversial part of any

[5] Or so they were in 1991 data. Revisiting the Maimonides' Rule research design with Israeli data for 2002–11, Angrist, Lavy, Leder-Luis, and Shany (2019) estimates class size effects tightly distributed around zero. Many countries have their own version of Maimonides' Rule, usually with cutoffs below 40. For example, Angrist, Battistin, and Vuri (2017) uses Italy's version to estimate causal effects of class size on the manipulation of standardized test scores. Sims (2008) uses Maimonides' Rule to document unintended consequences of a California class size reduction program that encouraged the use of classes mixing elementary school grades.

[6] In related work, Hoxby (2000) uses variation in the size of school-age populations in Connecticut school districts to construct instruments for class size.

IV story. My last example proposes and tests a strong but empirically compelling exclusion restriction: diversion from high-performing urban charter schools explains why enrollment at Chicago exam schools reduces student achievement. The lecture concludes with brief comments on the evolution of empirical economics since the 1980s.

2. EXAM TIME!

Suppose you'd like to run an RCT in which half of subjects are treated. You might randomly assign treatment by using your computer to draw a uniformly distributed random variable for each subject and treating those drawing values above one-half. This is the RCT version of RD: the running variable is uniformly distributed and the cutoff is one-half. Unlike the typical RD running variable, however, an RCT running variable is, by design, independent of subject characteristics and potential outcomes. You needn't worry about running variable control in this case.

Do comparisons above and below cutoffs—like the comparison of schools with 40 and 41 fourth graders behind Maimonides Rule estimates—really amount to something similar? Yes! Such comparisons exploit a feature of the physical world: provided the running variable has a continuous distribution, assignment rates approach the coin-toss rate of one-half when computed in a narrow window around the cutoff used to adjudicate treatment. In RD empirical work, the window around a cutoff is known as a *bandwidth*. Importantly, in the absence of any running variable manipulation that might interfere with continuity, the limiting probability of treatment as bandwidth shrinks is 0.5 for everybody, regardless of their characteristics or potential outcomes.

This remarkable fact can be seen in data on applicants to one of New York's highly coveted screened schools. By way of background, roughly 40% of New York City's middle and high schools select their applicants on the basis of test scores, grades, and other exacting criteria.[7] Only applicants ranked sufficiently highly are offered a screened-school seat. For screened-school applicants, running variables are the ranks schools assign their applicants. The cutoffs for screenedschool admission typically fall towards the top of the applicant distribution, rather than in the middle as in our hypothetical RCT.

Figure 2 documents the near random assignment of seats for a subset of applicants to New York's storied Townsend Harris high school (U.S. News and World Report recently ranked highly-selective Townsend Harris 12th nationwide, though New York has other even more selective schools). Bar height in the figure marks the *qualification rate*, that is, the likelihood of earning a Townsend Harris admissions score above that of the lowest-scoring applicant offered a seat. In our research on school assignment, my collaborators and I refer to qualification rather than admission because, in a centralized match such as that used by New York City high schools, qualification at Townsend Harris is necessary but not sufficient to be seated there.[8] The first pair of bars in Figure 2 show qualification rates *conditional* on a measure of pre-application achievement, labelled "baseline" in the figure. In particular, the bars mark qualification rates conditional on whether an applicant has upper-quartile or lower-quartile 6th grade scores.

Student achievement is highly persistent over time. Not surprisingly, therefore, Townsend Harris applicants with high baseline scores are much more likely to qualify

[7] More precisely, the 40% figure refers to school *programs*. New York's school buildings may host more than one program.

[8] Abdulkadiroğlu, Angrist, Narita, and Pathak (2017, 2022) derive the distribution of school assignments generated by the NYC high school match.

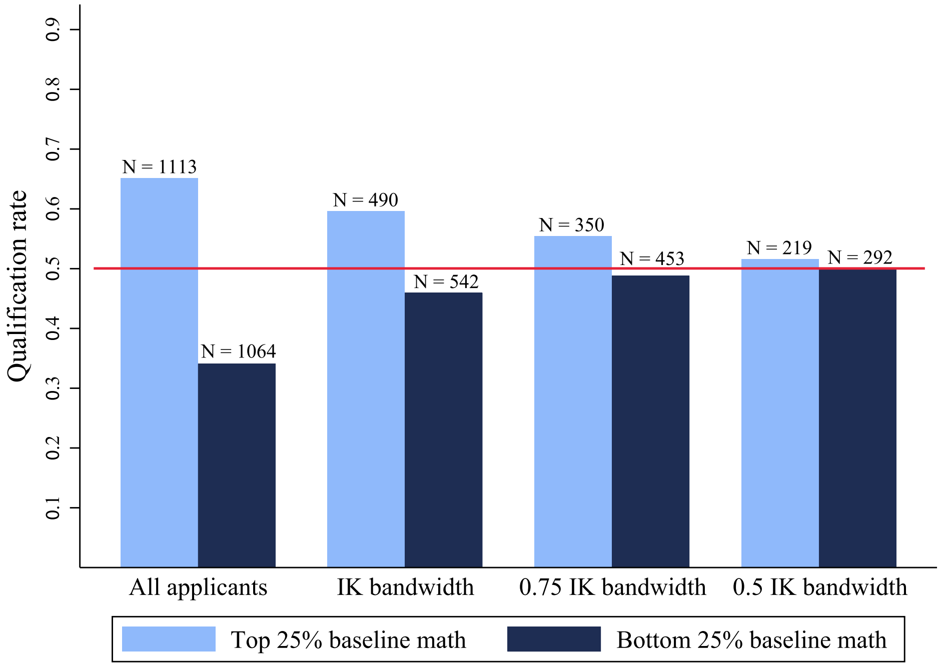


FIGURE 2.—Qualification rates near the Townsend Harris cutoff. *Notes*: This figure describes qualification rates for applicants to one of NYC's most selective screened schools, Townsend Harris (TH). The sample consists of applicants for 9th grade seats applying to TH in 2011–2013. The leftmost pair of bars compares all TH applicants whose baseline (6th grade math) scores fall in the upper and lower quartiles of the baseline score distribution. Other paired bars compare conditional qualification rates for applicants whose tie-breaker values lie within shrinking bandwidths around the TH cutoff. Bandwidths are estimated as suggested by Imbens and Kalyanaraman (2012), using a uniform kernel. Qualification is defined as clearing the relevant TH cutoff.

there than are applicants with low baseline scores. In a shrinking bandwidth around the school's cutoff, however, qualification rates in the two groups converge. The bar pair second from left shows conditional qualification rates in a window estimated as suggested by Imbens and Kalyanaraman (2012), the "IK bandwidth." Moving to the right, we see conditional qualification in a window of width 0.75 IK and then 0.5 IK. In the latter, the original sample size of about 2200 has fallen to around 500. Conditional qualification rates computed in the narrowest window are both remarkably close to one-half. This is what we'd expect to see were Townsend Harris to admit students by a coin toss rather than on the basis of test scores and grades. Yet, even when admissions operates by the latter rule, the data can be arranged so as to mimic the former.[9]

[9] The figure illustrates the following theorem. Suppose applicant i qualifies when continuous running variable R_i clears a fixed cutoff, τ, and assume the distribution of R_i has a continuously differentiable distribution function. Let $Q_i = 1[R_i > \tau]$ indicate qualification and let W_i be a random variable (like baseline scores) unchanged by qualification. Then,

$$\lim_{\delta \to 0} E\big[Q_i | W_i = w, R_i \in (\tau - \delta, \tau + \delta)\big] = 0.5.$$

Lee (2008) articulates the local randomization idea, but implements parametric RD. As far as I know, Cattaneo, Frandsen, and Titiunik (2015) is the first empirical application based on local randomization, using this idea for both estimation and inference. Frolich and Huber (2019) and Cattaneo, Titiunik, and Vazquez-Bare (2017) also discuss local randomization. Non-parametric RD using a shrinking data-driven bandwidth can be

The Elite Illusion

One of the most controversial questions I've studied is that of access to public exam schools like the Boston Latin School (America's first high school), Chicago's Payton and Northside selective enrollment high schools, and New York's legendary Brooklyn Tech, Bronx Science, and Stuyvesant specialized high schools, which have graduated 14 Nobel laureates between them (screened school Townsend Harris has graduated three Nobel-laureates, including economist Kenneth Arrow). Exam school proponents see these schools as democratizing public education. Wealthy families, they argue, can access advanced curricula in the private sector. Shouldn't bright low-income students be afforded the same chance? Critics of selective enrollment schools argue that, rather than expanding equity, exam schools are inherently biased against the Black and Hispanic students that make up the bulk of America's urban students. New York's unimaginably selective Stuyvesant, for example, admitted only 7 Black students to 9th grade in 2019, out of an incoming class of 895.

Motivated by the enduring controversy over selective admissions, my Blueprint Labs (https://blueprintlabs.mit.edu/) collaborators and I have examined the causal effects of exam school attendance in Boston, Chicago, and New York. This work has generated provocative findings, with profound implications for school assignment policy. Our first exam-school study, which looks at schools in Boston and New York, encapsulates these findings in the title, The Elite Illusion (Abdulkadiroğlu, Angrist, and Pathak (2014)). This alludes to the fact that, while exam school students undoubtedly have high test scores and other good outcomes, this success is not a *consequence* of exam school attendance. Our estimates consistently suggest that causal effects of exam school attendance on outcomes related to achievement and college attendance are zero, maybe even negative. The strong performance of exam school students is a manifestation of *selection bias* resulting from the process by which exam school students are chosen, rather than causal impact.

Data from Chicago's large exam school sector illustrate the elite illusion, while also laying the foundation for a causal story to which I'll return shortly (these data are analyzed in Abdulkadiroğlu, Angrist, Narita, Pathak, and Zárate (2017) and Angrist, Pathak, and Zárate (2019)).[10] The left panel of Figure 3 explains why exam schools are so attractive to parents. This panel plots peer mean achievement—specifically, the 8th grade test scores of an applicants' 9th grade classmates—against the admissions tie-breaker, for a subset of applicants to the group of nine Chicago exam schools open in 2009–12. Applicants rank up to six schools, while exam schools prioritize applicants using a common composite index formed from an admissions test, middle school GPA, and 7th grade standardized test scores. This composite tie-breaker is the running variable for an RD design that reveals what happens when an applicant is offered an exam school seat.

Because Chicago has many exam schools, the city uses a version of the celebrated Gale and Shapley (1962) deferred acceptance (DA) algorithm to adjudicate exam-school assignment (DA is celebrated in the 2012 Economics Nobel awarded to Alvin Roth and Lloyd Shapley). As it happens, the Chicago DA implementation is well-approximated by

justified by smoothness of conditional mean functions for potential outcomes rather than continuity of the running variable distribution, a distinction explored in Dong (2018) and Arai, Hsu, Kitagawa, Mourifié, and Wan (2022). Angrist and Rokkanen (2015) suggest conditioning on lagged outcomes to turn local random assignment into global random assignment in an RD context.

[10] Related Blueprint Labs exam-school research includes Idoux (2021), and Abdulkadiroğlu et al. (2022). Dobbie and Fryer (2014) and Barrow, Sartain, and de la Torre (2020) also use RD to study exam schools in New York and Chicago, respectively.

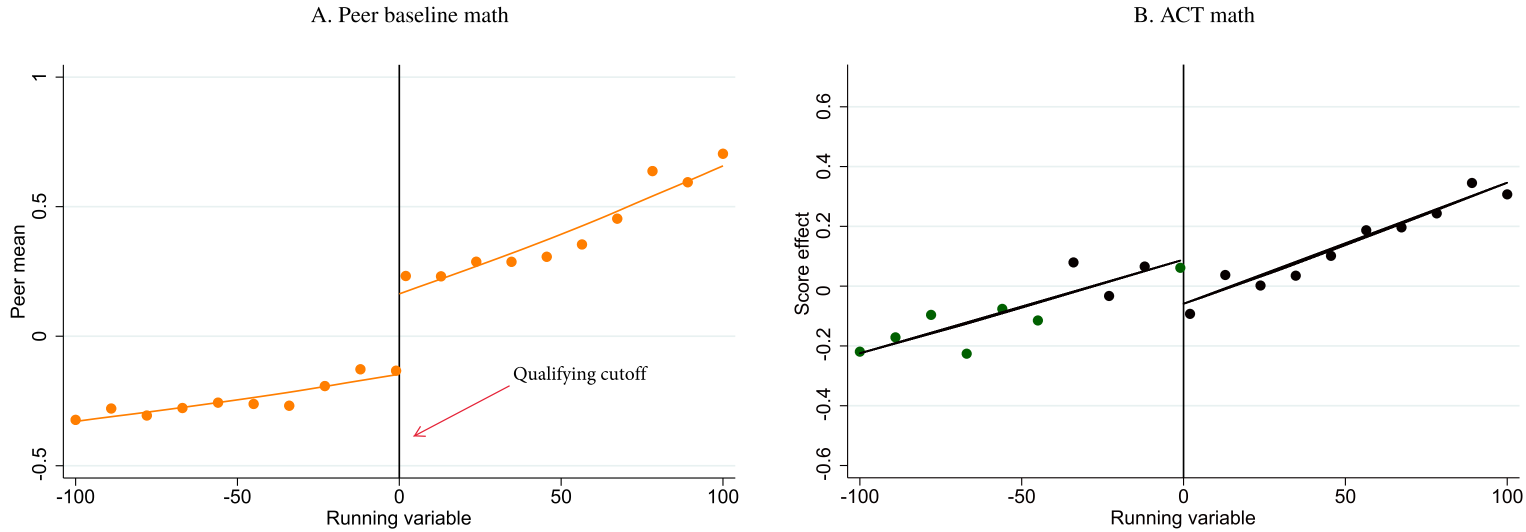


FIGURE 3.—Peer baseline achievement and ACT math effects at qualifying cutoffs for Chicago exam school applicants. *Notes*: This figure plots peer baseline math scores (Panel A) and ACT math scores (Panel B) against the exam school admissions composite tie-breaker. The sample consists of Chicago exam school applicants applying to at least one Noble charter school in the 2009–2012 application years. Baseline scores are taken from the 8th grade math Illinois Standards Achievement Test; ACT scores are from tests taken primarily in 11th grade. Baseline and ACT scores are standardized to have mean zero and unit standard deviation in the Chicago Public School district's test-taking population. A student's peers are all other 9th graders enrolled at the same school. The running variable is centered around the qualifying cutoff. Applicants who clear their qualifying cutoff are offered an exam school seat. Plotted points are averages in 10-unit windows; lines in the plots are estimated conditional mean functions smoothed using local linear regression, with a triangular kernel and bandwidth computed as suggested by Imbens and Kalyanaraman (2012). All variables are plotted after partialling out saturated qualifying-cutoff-by-tier-by-application-year fixed effects.

a simpler algorithm known colorfully as *serial dictatorship*. Under serial dictatorship, an exam school applicant is sure to be offered a seat somewhere when they clear the lowest cutoff in the set of cutoffs associated with the schools they rank. In the context of school assignment using serial dictatorship, we call this the qualifying cutoff.[11]

The left panel of Figure 3 shows a sharp jump in peer mean achievement for Chicago exam school applicants who clear their qualifying cutoff. This jump arises from the fact that most applicants offered an exam school seat take it. And applicants who enroll at one of Chicago's selective enrollment high schools are sure to be seated in 9th grade classrooms filled with academically precocious peers, since only the relatively precocious get in. The increase in peer achievement across the qualifying cutoff amounts to almost half of a standard deviation (the test scores used to measure peer quality have been scaled to have a mean of zero and a standard deviation of one in the district as a whole).

Precocious peers notwithstanding, the offer of an exam school seat does not appear to increase learning. The right-hand panel of Figure 3 plots applicants' ACT scores (a test taken mostly in 11th grade) against their tie-breaker values. This panel shows that exam-school applicants who clear their qualifying cutoff perform sharply *worse* on the ACT. Parents who enroll their children in one of Chicago's selective enrollment high schools in anticipation of accelerated learning are destined, on average, for disappointment.[12] What explains this? It takes a combination of IV and RD to untangle the forces behind this intriguing and unexpected negative impact. But first, some IV theory.

3. A LITTLE LATE

By connecting econometric ideas with the world of heterogeneous potential outcomes introduced in Rubin (1974, 1978), the LATE theorem offered a new understanding of empirical strategies involving IV and RD. The prize that Guido Imbens and I share recognizes the relevance of the LATE framework for modern empirical practice.

Guido and I overlapped for only one year at Harvard, where we had both signed on as assistant professors. In the fall of 1990, starting my second year on the job, I welcomed Guido to Cambridge with a pair of interesting instrumental variables. The first instrument, coded from draft lottery numbers randomly assigned in the 1970s, generates variation in Vietnam-era veteran status (Angrist (1990)). The second instrument, quarter of birth, arguably close to randomly assigned or at least serendipitous, interacts with compulsory attendance laws to generate variation in highest grade completed (Angrist and Krueger (1991)).

The draft lottery instrument relies on the fact that lottery numbers randomly assigned to birthdays determined Vietnam-era conscription risk. Even in the 1960s and 1970s, however, most American soldiers were volunteers, as all are today. The quarter-of-birth instrument uses the fact that men who are born earlier in the year typically start school

[11] Applicants who clear their qualifying cutoff are sure to be seated somewhere because at least one school judges their application acceptable. Depending on their tie-breaker value and preferences over schools, however, applicants may be offered a seat at a school they prefer to the school that determines their qualifying cutoff. The plots in Figure 3 were constructed by subtracting the qualifying cutoff from each applicant's admissions tie-breaker, so that all applicants face a common qualifying cutoff of zero.

[12] Barrow, Sartain, and de la Torre (2020) reports negative effects of Chicago exam-school offers on high school grades and the probability of attending a selective college. Dale and Krueger (2002) pioneered the study of the elite illusion in college, showing that college selectivity is unrelated to graduates' earnings, once account is taken of the schools to which students applied and were admitted. Mountjoy and Hickman (2020) apply this research design to large samples of public university applicants in Texas.

younger, and are therefore allowed to drop out of high school (on their 16th birthday) with less schooling completed than those born later. Yet, most people complete high school regardless of their quarter of birth. Guido and I soon began asking each other: What, really, did we learn from draft-lottery and quarter-of-birth instruments?

An early result in our quest for a new understanding of IV was a solution to the problem of selection bias in an RCT with partial compliance. Even in a randomized clinical trial, some assigned to treatment may choose to opt out, a fact that has long vexed trialists. Angrist and Imbens (1991) proved that in a randomized trial with partial compliance, the average causal effect of treatment on the treated is identified provided the control group has no access to treatment. This is in spite of the fact that those who comply with treatment in the treatment arm are likely to be a highly select group.

Unfortunately for us, we were late to the partial-compliance party. Not long after releasing our first coauthored working paper, we learned of Bloom (1984). The Bloom Result (as Steve Pischke and I called it in Angrist and Pischke (2009)) can be stated as follows. Consider a clinical trial that offers treatment randomly. Proportion π receive treatment when offered, while the rest opt out. Indicate those who are offered treatment with dummy variable Z_i, and those who take treatment with a dummy variable D_i. Denote potential outcomes for subject i in the treated and untreated states by Y_{1i} and Y_{0i}, respectively. The observed outcome is:

$$Y_i = Y_{0i} + D_i[Y_{1i} - Y_{0i}].$$

In other words, we see Y_{1i} for the treated and we see Y_{0i} for those not treated. $Y_{1i} - Y_{0i}$ is the causal effect of treatment on individual i, but this we can never see. We make do, therefore, with average treatment effects.

Bloom (1984) shows how to compute the average effect on the treated in this scenario. Let δ be the effect of treatment *assigned* on Y_i (trialists call this the *intention-to-treat effect* or ITT for short). Then,

$$E[Y_{1i} - Y_{0i}|D_i = 1] = \frac{\delta}{\pi}.$$

What could be simpler? This is the IV estimand that uses treatment assigned, Z_i, as an instrument for treatment received, D_i. To this day, I'm often asked how it can be true that in a scenario where subjects assigned to treatment selectively decline treatment, the average causal effect on the treated is knowable. Remarkably, Howard Bloom derived this result from first principles, making no connection to IV.

The LATE theorem (Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996)) generalizes Bloom's result. Maintaining the clinical trials analogy, let D_{1i} indicate subject i's treatment status when assigned to treatment and let D_{0i} indicate subject i's treatment status when assigned to control.[13] In addition to the assumptions underpinning Bloom, we added one more: assignment to treatment either leaves treatment status unaffected or makes it more likely (formally, $D_{1i} \geq D_{0i}$ for all i; the direction of the in-

[13] We owe this potential assignment notation to Gary Chamberlain. Writing me in November of 1991 with comments on an "early LATE draft," Gary noted that LATE as we had derived it had a "mysterious random variable in Condition 1." This was the error term we had used in a model of treatment assignment. Gary suggested we define D_{0i} and D_{1i} directly.

equality doesn't matter). Given this restriction, which we called *monotonicity*, LATE says:

$$E[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}] = \frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]} \tag{3}$$

$$= \frac{\delta}{\pi_1 - \pi_0}, \tag{4}$$

where π_1 and π_0 are compliance rates in the group assigned to treatment and the group assigned to control, respectively. The right-hand-side of (3) is again the IV estimand using treatment assigned as an instrument for treatment received. Motivated by Angrist and Krueger (1991), Angrist and Imbens (1995) extends LATE to ordered treatments like years of schooling, while Angrist, Graddy, and Imbens (2000) covers continuous treatments and simultaneous equations models.

3.1. *Ice Cream at Princeton, AIRtime at Harvard*

At Princeton and then Harvard, I read and reread Chamberlain (1984), Newey (1985), and Newey and West (1987). I was also lucky to be able to call on my Princeton Ph.D. advisor Whitney Newey and my Harvard colleague Gary Chamberlain in real life. Lengthy derivations begun in Whitney's office led often to Thomas Sweet's in Palmer Square, a reward for Whitney's patience. As a new assistant professor in 1990, I apprenticed to Gary by co-teaching his undergraduate econometrics course, an experience I've been building on ever since.

In Angrist (1990), my job market paper, I used draft-lottery dummies as instruments for veteran status in a two-sample linear IV procedure detailed in Angrist and Krueger (1992). Motivated by the fact that Hearst, Newman, and Hulley (1986) used the draft lottery to estimate veteran effects on mortality, I also began exploring IV methods for nonlinear qualitative response models. With little beyond bivariate probit to show for my efforts in this direction, Newey suggested I seek new causal knowledge from biostatistics maven Jamie Robins at the Harvard School of Public Health.[14] Robins advised me to abandon latent index models and turn instead to potential outcomes and the Rubin causal model. So I read and wrote Don Rubin.

Rubin's reply reached me in Jerusalem, where I had taken a position in Fall 1991. In the meantime, Guido found Don Rubin as well. It was Rubin who put us "on AIR," in Angrist, Imbens, and Rubin (1996), a follow-up to the 1994 paper, where, co-opting the Passover story, we redefined the four types of children (always-takers, never-takers, compliers, and defiers, described below). Along the way, we convinced Rubin of the utility of empirical strategies based on IV.

Convincing Don Rubin took some doing. His (September 1991) reply to me began: "Thanks for the copy of your paper on treatment effects ... I believe all the results, but I still cannot resonate to the approach." Among other complaints, Rubin wrote: "I don't know of any real success stories." I responded in October 1991, writing from Jerusalem: "I will try to explain why I find the IV framework so useful," going on to detail the draft lottery and the quarter-of-birth IV applications. Rubin replied with a longer letter that marked the beginning of our three-way collaboration. He agreed that the draft lottery

[14] Angrist (1991) shows that when the first stage causal effect of a dummy instrument on a dummy treatment is constant, the IV estimand using this instrument is an unconditional average treatment effect. This is implied by the LATE theorem because in this scenario $D_{1i} - D_{0i}$ is a constant. But I didn't know that at the time.

generates compelling instruments for Vietnam-era veteran status, but also wrote, "I want to make sure I really understand the assumptions you make without all the irrelevant linear model stuff."

And so on, back and forth. Along the way, Guido and I embraced the language of potential outcomes and eventually became fluent in it. But not right away: initially, Rubin and I argued every point. Then, in June 1992, I emailed Guido: "Never mind all my whining [about Rubin] from the previous email. I believe I've figured out how to link our earlier papers to 'the Rubin Way' ... the key is to follow up on something I think Don originally suggested: to define counterfactuals for the $2 * 2$ factorial experiment that manipulates *both* D and Z."

Double-indexing of potential outcomes allowed us to separate exclusion restrictions from independence assumptions, a feature integrated into the LATE framework by Angrist, Imbens, and Rubin (1996). The AIR paper characterizes exclusion by writing:

$$Y_i(d, z) = Y_i(d, z') \equiv Y_{di}, \tag{5}$$

where $Y_i(d, z)$ is the potential outcome seen when $D_i = d$ and $Z_i = z$. Independence asserts only that double-indexed potentials are independent of Z_i, while still allowing for possible direct effects of a candidate instrument, meaning $Y_i(d, z) \neq Y_i(d, z')$.

3.2. *LATE for Charter School*

The LATE theorem is formalized using the language of mathematical statistics. But the idea is pleasingly concrete and easy to grasp in practice. As in my undergraduate text with Steve Pischke (Angrist and Pischke (2014)), I'll explain the LATE idea here through a research question that has occupied me for almost two decades: the causal effect of charter school attendance on learning.[15]

Charter schools are public schools that operate independently of traditional American public school districts. A charter (the right to operate a public school) is typically granted for a limited period, subject to renewal conditional on good performance. Charter schools are free to structure their curriculum and school environment. Many charter schools extend instruction time by running long school days and continuing school on weekends and over the summer. The most controversial difference between charters and traditional public schools is the fact that the teachers and staff who work at the former rarely belong to labor unions. By contrast, most big-city public school teachers work under teachers' union contracts that closely regulate pay and working conditions.

The 2010 documentary film *Waiting for Superman* features schools belonging to the Knowledge is Power Program (KIPP). KIPP 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. The KIPP network serves a student body that is 95% Black and Hispanic,

[15] My interest in charter schools dates to 2003, when Michael Goldstein, then CEO of the Match Charter High School (https://www.matchschool.org/about/about-us/), invited Kevin Lang and me to use MATCH admissions lotteries to estimate causal effects of MATCH attendance. This initial effort failed to pan out because we were unable to obtain the needed data agreements. The first charter lottery analysis to which I contributed was released in 2009 (Abdulkadiroğlu et al. (2009)), later published as Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak (2011).

with over 80% of KIPP students poor enough to qualify for the federal government's subsidized lunch program.[16]

The American debate over education reform often focuses on the achievement gap, shorthand for large test score differences by race and ethnicity. Because of its focus on minority students, KIPP is often central in this debate, with supporters pointing to the fact that non-White KIPP students have markedly higher test scores than non-White students from nearby schools. KIPP skeptics have argued that KIPP's apparent success reflects the fact that KIPP attracts families with children who are more likely to succeed anyway.

A randomized trial might prove decisive in the debate over schools like KIPP. Luckily, while seats at KIPP are not filled by binding random assignment, there's a good deal of randomness in who gets one. This randomness comes from the fact that Massachusetts charter schools with more applicants than seats must offer their seats by lottery. Specifically, applicants are ordered according to a random lottery number, and offers made down this randomly ordered list until all available seats are taken.

A little over a decade ago my collaborators and I collected data on KIPP Lynn middle school admissions lotteries, thereby laying the foundation for charter school research published in Angrist, Dynarski, Kane, Pathak, and Walters (2010) and Angrist, Dynarski, Kane, Pathak, and Walters (2012). At the time, the KIPP middle school in Lynn, Massachusetts was the first school of its kind in New England. Some KIPP applicants bypass the lottery—those with previously enrolled siblings are guaranteed admission, while a few are categorically excluded (those too old for middle school, for example). Among the 371 applicants for 5th or 6th grade entry who were subject to random assignment in the four KIPP lotteries held from 2005–08 (and for whom we have post-application data on achievement), a total of 253 were offered a seat.

Perhaps surprisingly, a fair number of applicants offered a seat in the lottery failed to enroll come September. Some had moved away, while others ultimately preferred a traditional public school. Among applicants offered a seat, 199 (or about 79%) enrolled at KIPP the following school year. At the same time, 5 applicants (about 4.2%) not offered a seat nevertheless found their way into KIPP. The effect of an offer on KIPP enrollment rates is $\frac{199}{253} - \frac{5}{118} \approx 0.74$. In an IV analysis where offers are used as an instrumental variable for KIPP attendance, this 0.74 effect of offers on enrollment is the relevant *first stage*.

The IV empirical strategy sketched here looks at KIPP attendance effects on test scores for tests taken at the end of the grade following the application grade (these scores are from the end of 5th grade for those who applied in 4th and from the end of 6th grade for those who applied in 5th). As is common in research on student achievement, our score data are standardized by subtracting the mean and dividing by the standard deviation of scores in a reference population. In this case, the reference population contains all Massachusetts students in the relevant grade. Standardized scores are easily compared across populations and tests. As in many of Massachusetts' poorer cities and towns, average math scores in Lynn fall about three tenths of a standard deviation below the state mean (a gap that's written -0.3σ).

Among participants in KIPP entry lotteries, applicants offered a seat had standardized math scores close to zero (-0.003σ to be precise), that is, near the state mean. Because

[16] The case for No Excuses pedagogy begins with Martin Luther King Jr., who wrote in King (1967) that "Whatever pathology may exist in Negro families is far exceeded by this social pathology in the school system that refuses to accept a responsibility that no one else can bear and then scapegoats Negro families for failing to do the job." The first quantitative analysis of the No Excuses paradigm is likely Thernstrom and Thernstrom (2004).

KIPP applicants start with 4th grade scores that average roughly 0.3σ below the state mean, achievement at the level of the state average should be seen as impressive. By contrast, the average math score among those not offered a seat is about -0.358σ, a rssult typical of students residing in towns like Lynn.

Since lottery offers are randomly assigned, we can say with confidence that the offer of a seat at KIPP Lynn boosts math scores by an average of 0.355σ, a large effect that's also statistically precise. What does an *offer* effect around 0.36σ tell us about the effects of KIPP Lynn *attendance*? IV methods convert KIPP offer effects into KIPP attendance effects. In this case, the instrumental variable is a dummy variable that indicates KIPP applicants who receive offers. As in my discussion of RCTs, let Z_i denote this instrument. The causal effect of interest is that of D_i, a dummy indicating KIPP enrollment.

In general, three things are required of Z_i for it to be a valid instrument:

I. Z_i should have a causal effect on the treatment variable of interest, in this case KIPP enrollment, D_i. As noted above, this causal effect is called the *first stage*.

II. Z_i must also be randomly assigned or as good as randomly assigned, in the sense of being unrelated to potential outcomes, $Y_i(d, z)$, indexed by treatment and instrument. This is called the *independence assumption*.

III. Finally, IV logic requires an *exclusion restriction* postulating a single measured channel through which the instrument, Z_i, affects outcomes. This is expressed by equation (5). Here, the exclusion restriction amounts to the claim that the 0.355σ score differential between lottery winners and losers is entirely attributable to the 0.74 win-loss difference in KIPP attendance rates, that is, to the effect of Z_i on D_i.

An IV empirical strategy for charter school effects characterizes a chain reaction leading from the instrument to student achievement. The first link in this causal chain (the first stage) connects randomly assigned offers with KIPP attendance, while the second link (the one we're after) connects KIPP attendance with achievement. By virtue of the independence assumption and the exclusion restriction, the product of these two links generates the effect of offers on test scores:

$$\begin{aligned}&\text{Effect of offers on scores}\\&\quad=\{\text{Effect of offers on attendance}\}\times\{\text{Effect of attendance on scores}\}.\end{aligned}$$

The causal effect of KIPP *attendance* can therefore be written:

$$\text{Effect of attendance on scores} = \frac{\{\text{Effect of offers on scores}\}}{\{\text{Effect of offers on attendance}\}}.$$

This is a version of equation (4), expressed here in words.

The effect of the instrument (lottery offers) on outcomes (test scores) plays a central role in the IV story and therefore has a special name: this is the *reduced form*, denoted by δ in (4). Dividing the reduced form (0.355σ) by the first stage, the KIPP attendance effect is:

$$0.48\sigma \approx \frac{0.355\sigma}{0.745} = \frac{(-0.003\sigma)-(-0.358\sigma)}{0.787-0.042}.$$

Almost half a standard deviation gain in math scores is a remarkable result. Few education-related interventions have such large effects.[17]

[17] The full econometric analysis of KIPP is more involved than described here. Like many instrumental variables, the KIPP lottery offer instrument is valid only after conditioning on factors (like application year and

It's one thing to be able to compute an IV estimate and another to know what it means. Children may differ in the extent to which they benefit from KIPP. For some, perhaps a group that's highly motivated with a supportive family environment, the choice of KIPP over a traditional public school matters little; the causal effect of KIPP attendance on such applicants is zero. For others, KIPP attendance may matter greatly. LATE is an average of these different individual causal effects. In particular, LATE is an average causal effect for the population of children whose KIPP enrollment status is determined solely by the KIPP lottery.

As I've mentioned, the LATE idea is illuminated by the biblical story of Passover, which explains that there are four types of children, each with characteristic behaviors. Table I classifies applicants named Alvaro, Normando, and Camila, as a well as the fourth type of child, a *defier*. Applicant names hint at the way applicants *would respond* were they to win or lose the lottery. The columns in Table I indicate attendance choices made when $Z_i = 0$, while rows indicate choices made when $Z_i = 1$. The table covers all possible scenarios for every applicant, not just the scenarios we observe. In other words, the table records potential choices made when $Z_i = 1$, denoted D_{1i}, and potential choices made when $Z_i = 0$, denoted D_{0i}. Potential choices are like potential outcomes: for any given applicant, we see only one or the other.

Never-takers and *always-takers* are on the main diagonal: win or lose, their choice of school is unchanged. Always-takers like Alvaro are dying to go to KIPP; if they lose the KIPP lottery, their mothers find a way to enroll them in KIPP anyway, perhaps by re-applying. Never-takers like Normando worry about long school days and lots of homework. Normando doesn't really want to go to KIPP, and refuses to do so upon learning that he won the lottery. For Normando, $D_{1i} = D_{0i} = 0$, while, for Alvaro, $D_{1i} = D_{0i} = 1$. At the bottom left, *compliers* like Camila are happy to go to KIPP if they win a seat, while accepting the verdict if they lose. Camila complies with her lottery offer, attending KIPP when she wins but not otherwise. In other words, for Camila, $D_{1i} = 1$, $D_{0i} = 0$.

The term complier highlights the link between IV and the RCTs we aim to mimic with non-experimental empirical strategies. Many randomized trials randomize only the opportunity to be treated, while the decision to comply with the treatment protocol remains

TABLE I

THE FOUR TYPES OF CHILDREN.

		Lottery losers $Z_i = 0$	
		Doesn't attend KIPP $D_{0i} = 0$	Attends KIPP $D_{0i} = 1$
Lottery winners $Z_i = 1$	Doesn't attend KIPP $D_{1i} = 0$	Never-takers (*Normando*)	Defiers
	Attends KIPP $D_{1i} = 1$	Compliers (*Camila*)	Always-takers (*Alvaro*)

Note: KIPP = Knowledge Is Power Program.

entry grade) that determine the probability of being offered seat. Other controls, such as past achievement, are included to increase statistical precision. The complete analysis also allows for the fact that some children spend more time at KIPP than others between the time they apply and the time outcomes are measured. See Angrist et al. (2012) for details.

voluntary and non-random. RCT compliers are those who take treatment when the offer of treatment is made, but not otherwise. With lottery instruments, LATE is the effect of KIPP attendance on Camila and other compliers like her who enroll (take treatment) when offered a seat in the lottery, but not otherwise. IV methods are uninformative for Alvaro and Normando because lottery numbers are unrelated to their treatment status.

The defiers in Table I are those who enroll in KIPP only when *not* offered a seat in the lottery. Such perverse behavior makes IV estimates hard to interpret. With defiers as well as compliers in the data, the average effect of a KIPP offer can be zero even if everyone benefits from KIPP attendance. Luckily, defiant behavior is unlikely in charter lotteries and many other IV settings. We therefore assume such behavior is rare to nonexistent. This is the *monotonicity* assumption introduced in Imbens and Angrist (1994): the instrument is presumed to push affected applicants in one direction only.

The LATE theorem says that for any randomly assigned instrument with a non-zero first stage, satisfying both monotonicity and an exclusion restriction, the ratio of reduced form to first stage is the average causal effect of treatment on compliers. Each IV assumption plays a distinct role in establishing this: with no first stage, there's no charter experiment, while the independence assumption ensures the reduced form captures the causal effect of the instrument. The exclusion restrictions asserts that the reduced form is explained by KIPP attendance alone, while monotonicity plus exclusion are what make the KIPP attendance effect we seek proportional to the lottery-offer reduced form. These building blocks lead to a simple formula for causal effects on compliers.

The LATE framework is sometimes seen as limiting the relevance of econometric inference to a narrow set of context-specific effects. But the population of compliers is typically a group we'd very much like to learn about. In the KIPP example, compliers are children likely to be seated at KIPP were the school to expand and offer additional seats in a lottery. In Massachusetts, the number of charter seats is capped by law, so the consequences of charter expansion are central to state education policy (since the founding of Blueprint Labs, Massachusetts has seen two ballot initiatives on this matter). Cohodes, Setren, and Walters (2021) tackles the question of whether lottery-based estimates of charter effects predict learning gains when charter schools like KIPP are allowed to open new campuses and add seats. This investigation shows IV estimates using charter lotteries to offer a remarkably reliable guide to the performance of newly-opened charter campuses.

3.3. *No Excuses for not Closing the Achievement Gap*

The LATE theorem can be interpreted as saying that treatment is randomly assigned for compliers. LATE assumptions therefore identify the entire distribution of potential outcomes for compliers, as well as average treatment effects. To see this, suppose first that treatment, D_i, is randomly assigned in a stratified randomized trial, with strata encoded in covariate X_i. Conditional random assignment implies that:

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i | X_i. \tag{6}$$

Differences in treatment and control means within strata therefore yield conditional average causal effects:

$$\begin{aligned} E[Y_i|D_i = 1, X_i] - E[Y_i|D_i = 0, X_i] &= E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 0, X_i] \\ &= E[Y_{1i}|X_i] - E[Y_{0i}|X_i] \\ &= E[Y_{1i} - Y_{0i}|X_i]. \end{aligned} \tag{7}$$

Now, define $Y_i^*(c) \equiv 1(Y_i < c)$ for any constant, c. Because conditional indpendence holds for any function of potential outcomes, we may swap $Y_i^*(c)$ for Y_i in (7) to obtain:

$$E\big[Y_i^*(c)|D_i = 1, X_i\big] - E\big[Y_i^*(c)|D_i = 0, X_i\big] = \Pr[Y_{1i} < c|X_i] - \Pr[Y_{0i} < c|X_i].$$

The right-hand side of this expression is the difference in the distributions of Y_{1i} and Y_{0i} within strata, evaluated at c. Such distributional comparisons feature in RCTs evaluating life-saving vaccines and treatment regimens, where the distribution of interest is that of survival time. RCTs likewise reveal marginal distributions of potential outcomes ($\Pr[Y_{ji} < c]$; $j = 0, 1$), as well differences between them.

The LATE analog of the conditional independence expressed by (6) says that:

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i|D_{1i} > D_{0i}. \tag{8}$$

To see this, note first that $Z_i = D_i$ for compliers. Therefore,

$$E[D_i|Y_{1i}, Y_{0i}, D_{1i} > D_{0i}] = E[Z_i|Y_{1i}, Y_{0i}, D_{1i} > D_{0i}] = E[Z_i],$$

where the second equals sign uses the independence and exclusion assumptions. Expression (8) is remarkable because D_i itself is not randomly assigned. Yet, for compliers, D_i is independent of potential outcomes as if treatment were randomly assigned in an RCT.

Unsurprsingly, compliers are not labeled as such in any data set. Even so, a few simple formulas (based on Imbens and Rubin (1997) and developed further by my former Ph.D. student and MIT colleague Alberto Abadie) yield potential outcome distributions for the compliers in your data (Abadie (2002, 2003)).[18]

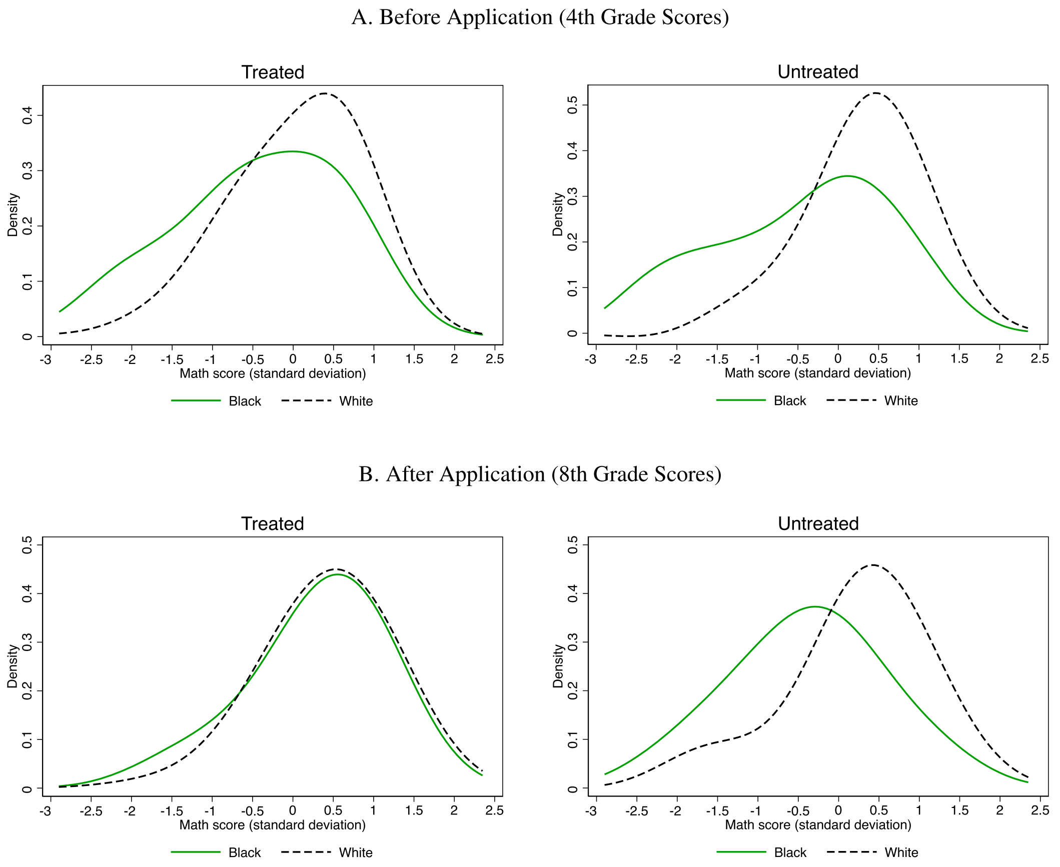
The importance of potential outcome distributions is illustrated by an analysis of charter school effects. Recall that our KIPP study was motivated in part by Black–White achievement gaps. Panel A in Figure 4 gives context for this concern by depicting the distribution of 4th grade scores for applicants to Boston charter middle schools. The two frames in Panel A show score distributions by race, tabulated separately for treated and untreated compliers. Treated compliers are compliers who attended a charter middle school, while untreated compliers did not. Because these are 4th grade scores, while middle school begins in 5th or 6th grade, the two sides of the figure are similar. In particular, both show score distributions for Black applicants shifted to the left of the corresponding score distributions for Whites.

By the end of 8th grade, the picture has changed markedly. This is documented in Panel B of Figure 4, which again shows score distributions by race, separately for treated and untreated compliers. Treated compliers finish middle school at a Boston charter in 8th grade. Like KIPP Lynn, Boston charter schools mostly employ No Excuses pedagogy. No Excuses charters boost achievement for most of their students, but those who enter the furthest behind tend to gain the most from charter attendance. Consequently, Black charter students, who start middle school with lower baseline scores, see their learning accelerated more by charters than do Whites (Angrist, Pathak, and Walters (2013), Chabrier,

[18] The cumulative distribution functions of Y_{ji}; $j = 0, 1$ are consistently estimated by the sample analog of

$$\Pr[Y_{ji} < c|D_{1i} > D_{0i}] = \frac{E\big[D_i^j(1 - D_i)^{1-j}Y_i^*(c)|Z_i = 1\big] - E\big[D_i^j(1 - D_i)^{1-j}Y_i^*(c)|Z_i = 0\big]}{(-1)^{1-j}\big(E[D_i|Z_i = 1] - E[D_i|Z_i = 0]\big)},$$

where, as before, $Y_i^*(c) \equiv 1(Y_i < c)$. Potential outcome densities can be obtained by replacing indicator functions with kernels; see Angrist, Cohodes, Dynarski, Pathak, and Walters (2016) and Abdulkadiroğlu, Pathak, and Walters (2018) for details.



A. Before Application (4th Grade Scores)

B. After Application (8th Grade Scores)

FIGURE 4.—Boston charter schools close the achievement gap. *Notes*: This figure depicts the distribution of math scores for treated charter-offer compliers, separately by race. Baseline (pre-application) scores are from 4th grade, while post-application scores are from 8th grade. The sample includes first-time applicants to seven Boston charter middle schools with 5th or 6th grade entry. These applicants were seeking seats in the 2005–2006 through 2008–2009 school years (see Walters (2018) for details). Complier distributions are estimated as described in Appendix A of Abdulkadiroğlu, Pathak, and Walters (2018).

Cohodes, and Oreopoulos (2016)). This differential impact is reflected in the bottom-left panel of the figure, which shows that, among treated compliers, the Black and White 8th grade score distributions have converged. Differences in 8th grade score distributions for untreated compliers, by contrast (shown at the bottom right of the figure), are changed little from baseline, with White students still clearly ahead of Black students.

3.4. *Where Do Babies Come From?*

IV estimation of the labor supply consequences of childbearing is motivated in part by the 20th century rise in married women's labor force participation, a trend that parallels declining marital fertility. Perhaps declining fertility explains increasing female labor supply. But the case for omitted variables bias in this context is clear: mothers with weak labor force attachment or low earnings potential may have more children than mothers with strong labor force attachment or high earnings potential. And causality might just as well run the other way, with increased female employment driving down fertility. This

TABLE II

IV ESTIMATES OF THE EFFECTS OF FAMILY SIZE ON LABOR SUPPLY.

			Twins Instrument		Same-Sex Instrument		Both
Dependent Variable	Mean	OLS (1)	First Stage (2)	IV Estimates (3)	First Stage (4)	IV Estimates (5)	2SLS Estimates (6)
Weeks worked	20.83	−8.98	0.603	−3.28	0.060	−6.36	−3.97
		(0.072)	(0.008)	(0.634)	(0.002)	(1.18)	(0.558)
	Overid: $\chi^2(1)$	—	—	—	—	—	5.3
	(p-value)						(0.02)
Employment	0.565	−0.176		−0.076		−0.132	−0.088
		(0.002)		(0.014)		(0.026)	(0.012)
	Overid: $\chi^2(1)$	—	—	—	—	—	3.5
	(p-value)						(0.06)

Note: This table reports OLS, IV, and 2SLS estimates of the effects of a third birth on labor supply using twins and sex composition instruments. Data are from the Angrist and Evans (1998) 1980 Census extract containing women aged 21–35 with at least two children. OLS models include controls for mother's age, age at first birth, ages of the first two children, and dummies for race. The sample size is 394,840.

makes the correlation between family size and mothers' employment or hours worked hard to interpret.

Bill Evans and I used two IV empirical strategies to overcome selection bias and capture causal effects of childbearing on parents' labor supply. The LATE theorem implies that these two instruments applied to the same causal relationship need not identify the same average causal effect. Different sets of compliers may be affected differently by the same intervention or treatment. Angrist and Evans (1998) and Angrist and Fernández-Val (2013) show that this is more than a theoretical possibility. Causal effects of childbearing depend in part on where the babies in question come from.

The first Angrist and Evans (1998) fertility instrument indicates the occurrence of twins at second birth in samples of mothers with at least two children (Rosenzweig and Wolpin (1980) is the first study to use twin births to instrument family size). The second instrument, also coded for women who have had at least two children, indicates whether first- and second-born children are of the same sex. American parents show little preference for boys or girls (the probability of having a second birth is similar whether the first-born is male or female). But they do seek a diversified sibling-sex portfolio: in families where the first and second-born children are both boys or both girls, the likelihood of a third birth jumps.

The twins first stage in 1980 Census data is about 0.6, an estimate reported in column 2 of Table II (adapted from Angrist and Fernández-Val (2013)). This means that 40 percent of mothers with two or more children would have had a third birth without twinning, while a multiple second birth increases this proportion to 100 percent. Validity of the twins instrument rests on the claim that multiple births are unrelated to potential outcomes indexed against childbearing and multiple birth events, and that a multiple birth affects labor supply solely by increasing fertility.[19]

[19] These conditions are unlikely to be met in a contemporary sample because the twin birth rate is boosted by in-vitro fertilization (IVF) and related fertility treatments. IVF is now used much more widely than in 1980, and is more common among older and more educated women.

The same-sex first stage is an order of magnitude smaller than the twins first stage. Parents of a same-sex sibship are about six percentage points more likely to have a third child than are parents of a mixed-sex sibship. This is documented in column 4 of Table II (38% of mixed-sex parents have a third child). Validity of the same-sex instrument rests on the claim that sibling sex composition is essentially random and affects mothers' labor supply solely by increasing fertility.

Twins-IV estimates suggest that a third-birth reduces mothers' weeks worked by a little over 3 weeks, with an employment reduction of about 0.08 points. These results, shown in column 3 of Table II, are smaller in absolute value than the corresponding ordinary least squares (OLS) estimates reported in the first column of the table. The latter, computing using a set of controls listed below the table, suggest a third birth reduces mothers' employment rates by around 18 percentage points, accompanied by 9 fewer weeks of work. In view of the twins-IV estimates, however, these large OLS estimates are almost certainly exaggerated by selection bias.

IV estimates constructed using the same-sex instrument, reported in column 5 of Table II, are substantially more negative than the corresponding twins-IV estimates (though still much smaller than OLS). Perhaps the gap between the two sets of IV estimates is a chance finding, due to sampling variance in the estimates. The last column of Table II reports 2SLS estimates of third-birth effects computed using twins and same-sex instruments together, along with the associated over-identification test statistic, which implicitly tests the null hypothesis that the underlying one-instrument-at-a-time IV estimates capture the same causal effect. This test generates p-values of 0.02 and 0.06, implying that the twins and same-sex IV estimates are statistically distinguishable, that is, differences between them are unlikely to be due to chance alone.[20]

In Angrist and Fernández-Val (2013), my former Ph.D. student Iván Fernández-Val and I argue that differences between twins and same-sex IV estimates reflect differences in the populations of twins and same-sex compliers. Since all mothers of second-born twins have at least three children, there are no twins never-takers. LATE logic therefore implies that twins instruments identify the average effect of a third child on *all* women who choose to have only two. Formally, since $D_{1i} = 1$ for all i,

$$\begin{aligned} E[Y_{1i} - Y_{0i} | D_{1i} = 1, D_{0i} = 0] &= E[Y_{1i} - Y_{0i} | D_{0i} = 0] \\ &= E[Y_{1i} - Y_{0i} | D_{0i} = 0, Z_i = 0] \\ &= E[Y_{1i} - Y_{0i} | D_i = 0]. \end{aligned}$$

In other words, twins instruments reveal the effect of a third birth on women who prefer two children only (the second equals sign above uses independence and exclusion; the third uses the fact that, for a twins instrument, $D_i = 0$ if and only if $D_{0i} = Z_i = 0$). The same-sex instrument, by contrast, captures childbearing effects on women who can be nudged into additional childbearing by the desire for a mixed-sex sibship.

Why are differences between twins and same-sex compliers economically important? In the Gronau (1977) model of labor supply, services like child care can be purchased in the market or provided in the home. The choice between these options is determined by

[20] 2SLS efficiently combines multiple instruments by using the fitted values generated by a first-stage equation with all instruments included on the right hand side as a single combined instrument. Models with more than one instrument for a single causal effect are said to be over-identified. The over-identification test statistic is proportional to the R-squared from a regression of 2SLS residuals on the instruments and covariates included in the first stage. See Hausman (1983) for details.

a mother's market wage. Consistent with the fact that they choose smaller families, twins compliers are especially likely to be college-educated, while college education and the higher wages education brings encourage out-of-home child care. This facilitates mothers' labor force participation in the wake of a third birth. Same-sex compliers, by contrast, are about two-thirds as likely as the typical mother of two to have a college degree, and are therefore more likely than twins compliers to use home child care in response to a third birth. Reliance on home care boosts the work-limiting consequences of childbirth.[21]

This tale of two instruments shows how the LATE theorem interacts with economic theory to reconcile contrasting results from two natural experiments when these experiments identify features of the same underlying causal relationship. The LATE framework allows us to describe distinct complier characteristics. Differences in complier characteristics linked with fundamental economic forces then suggest a theoretical explanation for differences in LATEs.

4. CONSTRUCTING CAUSAL STORIES

My Blueprint Labs colleagues and I have investigated many intriguing causal stories. One of the most surprising, told through a combination of IV and RD, resolves the puzzle of negative Chicago exam school effects. As a reminder, the challenge I set out in Section 2 is to explain why offers of a seat at one of the Windy City's coveted selective enrollment high schools appear to reduce learning rather than increase it.

Economic reasoning highlights alternatives. What's the alternative to an exam school? For most applicants to Chicago exam schools, the leading non-exam alternative is a traditional public school. But many of Chicago's rejected exam-school applicants enroll in charter schools. The offer of an exam-school seat therefore reduces the likelihood of charter-school attendance. Specifically, exam-school offers divert applicants away from high schools in the Noble Network of charter schools. Noble, deploying pedagogy much like KIPP's, is one of Chicago's most visible charter providers, enrolling roughly 40% of the city's 9th grade charter students.

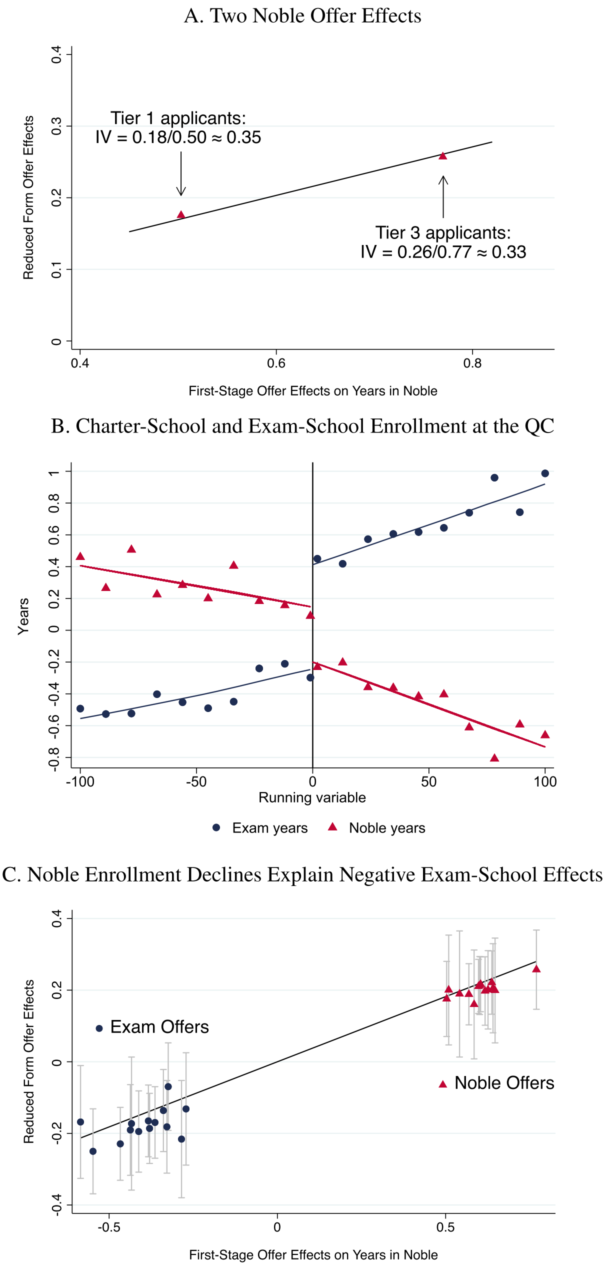
Also like KIPP, convincing evidence on Noble effectiveness comes from admissions lotteries: when their campuses are over-subscribed, Noble schools offer seats by random assignment. Noble applicants seated at Noble schools as a result of these admissions lotteries have higher ACT scores as a result (Davis and Heller (2019) is the first study using lotteries to document Noble effectiveness).

This evidence of Noble impact can be seen in Panel A of Figure 5.[22] The x-axis shows effects of Noble lottery offers on years enrolled at Noble; this is the first stage for an IV estimator using a dummy indicating Noble lottery offers as an instrument for Noble enrollment (I switch here to years enrolled rather than a dummy indicating any charter attendance because the time Noble students spend at Noble ahead of their ACT tests varies from one student to another). Panel A has another feature that distinguishes it from the simpler KIPP analysis: this plot shows estimates for two groups, one for Noble applicants who live in Chicago's lowest-income neighborhoods (labeled "Tier 1") and one for Noble applicants who live in higher-income areas ("Tier 3").[23]

[21] College graduation rates among compliers can be computed and compared using the fact that the probability compliers have Bernoulli characteristic $x_i = 1$, relative to the marginal probability that $x_i = 1$, is given by the ratio of the first stage conditional on $x_i = 1$ to the unconditional first stage. See Angrist and Pischke (2009) for details.

[22] Like Figure 3, this figure is derived from exhibits in Angrist, Pathak, and Zárate (2019).

[23] Most Chicago public school students are low-income; tiers classify relative income within the city.



A. Two Noble Offer Effects

B. Charter-School and Exam-School Enrollment at the QC

C. Noble Enrollment Declines Explain Negative Exam-School Effects

FIGURE 5.—Diversion from charters explains negative Chicago exam school effects. *Notes*: Panel A plots Noble offer effects for the Tier 1 and Tier 3 applicant groups. Panel B plots exam and Noble enrollment rates against the exam school admissions composite score; the vertical line here marks the qualifying cutoff (as in Figure 3). Panel C is a visual instrumental variable (VIV) plot of exam and Noble offer effects for a set of 14 covariate-defined groups. Covariate-specific estimates are computed one at a time in the relevant subsamples. The slope of the line through the estimates in Panel A rounds to 0.34; the slope in Panel C rounds to 0.36. The lines in Panels A and C are forced to pass through the origin. Whiskers in Panel C mark 95% confidence intervals for each reduced form estimate.

Recall the IV chain reaction: the reduced-form effect of an instrument on outcomes equals the causal effect of interest times the corresponding first stage. Each point in Panel A of Figure 5, which has coordinates given by (first stage, reduced form), therefore implies an IV estimate. In this case, we have:

$$\begin{aligned}&\text{Effect of Noble enrollment on ACT scores}\\&\quad=\frac{\{\text{Effect of Noble offers on ACT scores}\}}{\{\text{Effect of Noble offers on Noble enrollment}\}}.\end{aligned}$$

For Tier 1 applicants, this IV estimate is (after rounding) 0.35, while for Tier 3 the IV estimate rounds to 0.33.

For Noble applicants from both tiers, first-stage and reduced-form estimates imply an impressive yearly Noble enrollment impact of about a third of a standard deviation. A line drawn through these two points—plus the origin—fits well (the line is depicted in the figure, but the origin is not). Moreover, since the fitted line has an intercept of zero, its slope ("rise over run") is given by the two IV estimates that lie on it (empirically, the slope of the line comes out in-between the two estimates, at 0.34σ). The fact that the line connecting IV estimates for different groups runs through the origin supports an exclusion restriction which says that, in a group for which Noble offers are unrelated to Noble enrollment, we should see no reduced-form effect of Noble offers on test scores.

How consistent is the evidence for a Noble enrollment effect on the order of 0.34σ per year? The triangles plotted in the upper right area of Panel C of Figure 5 show first-stage and reduced-form Noble offer effects for 14 groups (2 more tiers and 10 groups defined by demographic characteristics related to race, sex, family income, and baseline scores). Although not a perfect fit, these points cluster around a line with slope 0.36σ, close to the slope of the line in Panel A. Again, consistent with an exclusion restriction that attributes reduced-form effects of Noble offers on test scores to first-stage effects of Noble offers on Noble enrollment, the fitted line passes through the origin.

The fact that the line in Panel C fits reasonably well bears a digression. As noted in the discussion of twins and same-sex instruments in Section 3.4, over-identification tests compare alternative IV estimates of the same causal effect. In a constant-effects causal framework, alternative IV estimates of the Noble enrollment effect should be similar unless one of the instruments is invalid. But LATEs using different instruments can differ even when all instruments are valid. Even in the LATE framework, however, the fact that the reduced form is proportional to the corresponding first stage has testable implications. In particular, reduced-form effects associated with a particular first stage should not be implausibly large, and reduced-form effects of instruments for which the first stage is zero should be zero too. These restrictions hold even in the absence of constant causal effects, though constant effects is a simple way to motivate them.[24]

What do the Noble IV estimates in Panels A and C of Figure 5 have to do with effects of *exam-schools*? The answer appears in Panel B of Figure 5, which complements the RD plots in Figure 3 with an added twist. The line fit to the dots in Panel B shows, as we should expect, that exam school enrollment jumps for applicants who clear their qualifying cutoff

[24] Building on Balke and Pearl (1997) and Imbens and Rubin (1997), LATE-compatible tests of instrument validity are developed in Heckman and Vytlacil (2005), Kitagawa (2015), Huber and Mellace (2015), and Frandsen, Lefgren, and Leslie (2019). Angrist, Lavy, and Schlosser (2010) uses the "no first stage, no reduced form" restriction to assess the validity of twins and same-sex instruments for family size. Sun and Wüthrich (2022) introduces a class of IV estimators that imposes these restrictions.

(qualifying applicants are offered an exam-school seat somewhere). Specifically, qualification boosts years enrolled at an exam school by 0.61. At the same time, the line fit to the triangles in Panel B shows that exam-school qualification reduces years of *Noble* enrollment by 0.37. This is the diversionary impact of exam school offers on Noble enrollment.

IV allows us to go out on a limb with strong and potentially falsifiable claims regarding the mechanism underlying a particular set of causal effects. Here's a strong causal claim: the primary force driving the reduced-form impact of exam-school qualification on ACT scores is the diversion away from charters seen in Panel B of Figure 5, that is, the negative impact exam-school offers have on Noble enrollment. In this account of exam-school offer reduced forms, exam-school offers leave achievement unchanged for exam-school applicants not diverted from Noble.

In support of this claim, note the points plotted as dots in Panel C of Figure 5, which lie well to the left of zero on the x-axis. The x-coordinates for these points mark the effect of exam-school qualification on Noble enrollment for various groups of applicants. Because exam-school offers divert many exam-school applicants away from Noble, these estimates are negative (as with the triangles for Noble offers, there's a dot for each of 14 covariate groups).

We've already seen that Noble applicants offered a seat in a Noble admissions lottery realize large ACT math score gains as a result. With this in mind, consider *exam-school offers* as an instrument for *Noble enrollment*. If exam school qualification reduces time at Noble by 0.37 years, and each year of Noble enrollment boosts ACT math scores by about 0.36 standard deviations, as suggested by the line plotted in Panel C of Figure 5, we should expect reduced-form effects of exam school qualification to be about -0.13σ. This is roughly consistent with the set of reduced-form exam-qualification estimates plotted as dots at the bottom left of Panel C (again, reduced form estimates are on the y-axis). The fit isn't perfect; reduced-form estimates in this part of the figure cluster closer to -0.15 than to -0.13, but this small shortfall can be attributed to sampling variance. The reduced form estimates also vary roughly in proportion to the group-specific first-stages effects of exam school offers on Noble enrollment.

The causal story told here postulates diversion away from charter schools as the primary mechanism by which Chicago exam school offers affect achievement. In other words, it's Noble enrollment that generates an exclusion restriction when we use exam-school offers as an instrument. Because the line drawn in Panel C runs through the origin, this exclusion restriction commits us to the diversion hypothesis: in applicant groups where exam-school offers have little or no effect on charter-school enrollment, these offers are predicted to leave ACT scores unchanged. The reduced-form and first-stage estimates plotted in the figure need not have aligned with this. It's revealing to know that they do.

5. EMPIRICAL ECONOMICS GETS SERIOUS

I computed the IV estimates in my Princeton Ph.D. thesis on a mainframe computer using 9-track tape and leased storage space on a crowded communal hard drive. Princeton graduate students mastered IBM job control language (https://en.wikipedia.org/wiki/Job_Control_Language), the better to manipulate tape reels the size of a cheesecake (https://en.wikipedia.org/wiki/9-track_tape) (overwrite your tape in haste, repent at leisure). Thankfully, empirical work today is less labor-intensive.

What else has improved in the modern empirical era? In Angrist and Pischke (2010), Steve Pischke and I coined the phrase "credibility revolution." By this, we meant applied microeconomists' growing use of transparent empirical strategies designed to answer specific causal questions. Previously, econometric analysis aimed mostly to estimate parameters governing the behavior of an economic model, often without a particular causal question in mind. Earlier work also paid little attention to the sources of variation underpinning a particular set of estimates. Empirical strategies emphasizing specific questions and the research designs deployed to answer them have since yielded a steady flow of credible causal conclusions.[25]

The rise of credible applied econometrics fueled a wave of theoretical econometric innovation that continues today. Much of the design-focused methodological agenda builds on Rosenbaum and Rubin's (1983) propensity score theorem. This theorem contributed to the credibility revolution by shifting econometricans' attention to the process determining treatment assignment rather than on models for outcomes. Dehejia and Wahba (1999) was the first econometric study to demonstrate the value of the propensity score for applied work, while Hahn (1998), Hirano, Imbens, and Ridder (2003), and Angrist and Hahn (2004) address important theoretical questions about the score. Angrist and Kuersteiner (2011) and Angrist, Jordà, and Kuersteiner (2018) contribute to this agenda by extending the propensity score/potential outcomes framework to causal questions in macroeconomics.

More recently, Belloni, Chernozhukov, and Hansen (2014) and Chernozhukov et al. (2018) use machine learning to model the propensity score while also modeling outcomes. These contributions can be seen as extending the Robins (2000) notion of double robustness to a wider class of empirical strategies. In principle, lasso and related ML tools offer a data-driven, theoretically-principled scheme to select from among many possible control variables. At the same time, Wüthrich and Zhu (2021) argue that in regression applications where the number of candidate controls is high but below sample size, OLS with all controls works well (though in some cases, the good performance of OLS with many controls requires use of the many-covariate-robust standard errors introduced by Cattaneo, Jansson, and Newey (2018)). ML for instrument selection, implemented as suggested by Belloni, Chen, Chernozhukov, and Hansen (2012), seems ill-suited for my type of IV application (Angrist and Frandsen (2022) gives evidence on this point).

A distinctive RD methodology continues to bloom. In a cascade of contributions, econometricians have tackled the vexing details of nonparametric bandwidth choice for nonparametric RD (as in Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, Farrell, and Titiunik (2017)). Nonparametric RD also requires a modicum of continuity, though Kolesár and Rothe (2018) shows how we can sometimes make do with a discrete running variable. De Chaisemartin and Behaghel (2020) solves estimation problems arising in RD designs relying on behaviorally-determined cutoffs, as is the case with the RD designs used in our lab's work on schools.

The outsized role played by IV in modern empirical work has prompted an explosion of research into the finite-sample behavior of IV estimators. Influential contributions like Staiger and Stock (1997), Stock and Yogo (2005), and Moreira (2003) (to name a few) are motivated by the Bound, Jaeger, and Baker (1995) critique of the heavily over-identified models used in Angrist and Krueger (1991). Research on the finite sample behavior of IV estimators is recently summarized in Andrews, Stock, and Sun (2019). My view, at odds with many theorists', is that old-fashioned asymptotic inference is often satisfactory. In Angrist and Kolesár (2021), Michal Kolesár and I argue that, when it comes to just-identified IV at least, worries about weak instruments are overblown.

[25] Card (2022) traces the intellectual roots of empirical econometrics' shift towards strong research designs.

I'm looking forward to solutions to the many problems my labmates and I encounter in our empirical work on causal effects. These include the development of new tools for estimation and inference in empirical strategies combining research design with market design (key identification results appear in (Abdulkadiroğlu et al. (2017, 2022))). Inference with clustered data remains as vexing as ever, though Abadie, Athey, Imbens, and Wooldridge (2022) makes the clustering issue easier to manage. RD is not foolproof: working on Angrist et al. (2019), I learned that school enrollment is an easily-manipulated running variable. More and better solutions for this sort of problem, as in Gerard, Rokkanen, and Rothe (2020), would be welcome.

A few notes in a minor key: empirical economics is more exciting and relevant than ever, but undergraduate econometric instruction has been slow to fully embrace modern empirical strategies. Angrist and Pischke (2017) argues that compelling empirical applications are the way forward in the classroom. In the domain of research on schools, I worry that hostility to standardized testing may cripple the measurement of school effectiveness (Olson and Jerald (2020) documents anti-testing trends). My labmates and I aspire to measure school quality fairly. Recently, for example, we've shown how to mitigate racial bias and the elite illusion in school ratings (Angrist, Hull, Pathak, and Walters (2017, 2021a,b)). Yet, without assessing their reading skills, how are we to know whether children are learning to read?

I'll conclude by saying that I'm proud to be part of the contemporary empirical economics enterprise and gratified beyond words to be recognized for contributing to it. Back at Princeton in the late 1980s, my graduate classmates and I chuckled reading Leamer's (1983) lament that "no economist takes another economist's empirical work seriously." This is no longer true. Empirical work today aspires to craft convincing causal stories. Not that every effort succeeds, far from it. But, as any economics job market candidate will tell you, empirical work carefully executed and clearly explained is taken seriously indeed. I hope that today's Ph.D. students will join me in seeing this as a measure of our enterprise's success.

REFERENCES

ABADIE, ALBERTO (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292. [2525]

——— (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [2525]

ABADIE, ALBERTO, SUSAN ATHEY, GUIDO W. IMBENS, AND JEFFREY M. WOOLDRIDGE (2022): "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics* (forthcoming). [2534]

ABDULKADIROĞLU, ATILA, JOSH D. ANGRIST, SARAH R. COHODES, SUSAN M. DYNARSKI, JON FULLERTON, THOMAS KANE, AND PARAG A. PATHAK (2009): "Informing the Debate: Comparing Boston's Charter, Pilot and Traditional Schools," The Boston Foundation. [2520]

ABDULKADIROĞLU, ATILA, JOSHUA D. ANGRIST, SUSAN M. DYNARSKI, THOMAS J. KANE, AND PARAG A. PATHAK (2011): "Accountability and Flexibility in Public Schools: Evidence From Boston's Charters and Pilots," *Quarterly Journal of Economics*, 126, 699–748. [2520]

ABDULKADIROĞLU, ATILA, JOSHUA D. ANGRIST, YUSUKE NARITA, AND PARAG A. PATHAK (2017): "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 85, 1373–1432. [2513,2534]

——— (2022): "Breaking Ties: Regression Discontinuity Design Meets Market Design," *Econometrica*, 90, 117–151. [2513,2515,2534]

ABDULKADIROĞLU, ATILA, JOSHUA D. ANGRIST, YUSUKE NARITA, PARAG A. PATHAK, AND ROMAN A. ZÁRATE (2017): "Regression Discontinuity in Serial Dictatorship: Achievement Effects at Chicago's Exam Schools," *American Economic Review: Papers & Proceedings*, 107, 240–245. [2515]

ABDULKADIROĞLU, ATILA, JOSHUA D. ANGRIST, AND PARAG A. PATHAK (2014): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82, 137–196. [2515]

ABDULKADIROĞLU, ATILA, PARAG A. PATHAK, AND CHRISTOPHER R. WALTERS (2018): "Free to Choose: Can School Choice Reduce Student Achievement?" *American Economic Journal: Applied Economics*, 10, 175–206. [2525,2526]

ANDREWS, ISAIAH, JAMES H. STOCK, AND LIYANG SUN (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753. [2533]

ANGRIST, JOSHUA D., AND JINYONG HAHN (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *The Review of Economics and Statistics*, 86, 58–72. [2533]

ANGRIST, JOSHUA, AND MICHAL KOLESÁR (2021): "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV," NBER Working Paper No. 29417, October. [2533]

ANGRIST, JOSHUA D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records," *American Economic Review*, 80, 313–336. [2517,2519]

——— (1991): "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," NBER Working Paper No. 0115, November. [2519]

ANGRIST, JOSHUA D., AND WILLIAM N. EVANS (1998): "Children and Their Parents' Labor Supply: Evidence From Exogenous Variation in Family Size," *American Economic Review*, 88, 450–477. [2527]

ANGRIST, JOSHUA D., AND IVÁN FERNÁNDEZ-VAL (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," in *Advances in Economics and Econometrics*. Econometric Society Monographs, Econometrics, Vol. III, ed. by Daron Acemoglu, Manuel Arellano, and Eddie Dekel. Cambridge University Press, Chapter 11, 401–434. [2527,2528]

ANGRIST, JOSHUA D., AND BRIGHAM FRANDSEN (2022): "Machine Labor," *Journal of Labor Economics*, 40, S97–S140. [2533]

ANGRIST, JOSHUA D., AND GUIDO W. IMBENS (1991): "Sources of Identifying Information in Evaluation Models," NBER Working Paper No. 0117, December. [2518]

——— (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. [2519]

ANGRIST, JOSHUA D., AND ALAN B. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014. [2517,2519,2533]

——— (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables With Moments From Two Samples," *Journal of the American Statistical Association*, 87, 328–336. [2519]

ANGRIST, JOSHUA D., AND GUIDO M. KUERSTEINER (2011): "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score," *The Review of Economics and Statistics*, 93, 725–747. [2533]

ANGRIST, JOSHUA D., AND VICTOR LAVY (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575. [2511,2512]

ANGRIST, JOSHUA D., AND JORN-STEFFEN PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. [2511,2518,2529]

ANGRIST, JOSHUA D., AND JÖRN-STEFFEN PISCHKE (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics," *Journal of Economic Perspectives*, 24, 3–30. [2532]

——— (2014): *Mastering 'Metrics: The Path From Cause to Effect*. Princeton: Princeton University Press. [2520]

——— (2017): "Undergraduate Econometrics Instruction: Through Our Classes, Darkly," *Journal of Economic Perspectives*, 31, 125–144. [2534]

ANGRIST, JOSHUA D., AND MIIKKA ROKKANEN (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff," *Journal of the American Statistical Association*, 110, 1331–1344. [2515]

ANGRIST, JOSHUA D., ERICH BATTISTIN, AND DANIELA VURI (2017): "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics*, 9, 216–249. [2512]

ANGRIST, JOSHUA D., SARAH R. COHODES, SUSAN M. DYNARSKI, PARAG A. PATHAK, AND CHRISTOPHER R. WALTERS (2016): "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry and Choice," *Journal of Labor Economics*, 34, 275–318. [2525]

ANGRIST, JOSHUA D., SUSAN M. DYNARSKI, THOMAS J. KANE, PARAG A. PATHAK, AND CHRISTOPHER R. WALTERS (2010): "Inputs and Impacts in Charter Schools: KIPP Lynn," *American Economic Review: Papers & Proceedings*, 100, 239–243. [2521]

——— (2012): "Who Benefits From KIPP?" *Journal of Policy Analysis and Management*, 31, 837–860. [2521, 2523]

ANGRIST, JOSHUA D., KATHRYN GRADDY, AND GUIDO W. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models With an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527. [2519]

ANGRIST, JOSHUA D., PETER D. HULL, PARAG A. PATHAK, AND CHRISTOPHER R. WALTERS (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 132, 871–919. [2534]

——— (2021a): "Credible School Value-Added With Undersubscribed School Lotteries," MIT Blueprint Labs Working Paper. [2534]

——— (2021b): "Race and the Mismeasure of School Quality," NBER Working Paper No. 29608, December. [2534]

ANGRIST, JOSHUA D., GUIDO W. IMBENS, AND DONALD B. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [2518-2520]

ANGRIST, JOSHUA D., VICTOR LAVY, JETSON LEDER-LUIS, AND ADI SHANY (2019): "Maimonides' Rule Redux," *American Economic Review: Insights*, 1, 309–324. [2512,2534]

ANGRIST, JOSHUA D., VICTOR LAVY, AND ANALIA SCHLOSSER (2010): "Multiple Experiments for the Causal Link Between the Quantity and Quality of Children," *Journal of Labor Economics*, 28, 773–824. [2531]

ANGRIST, JOSHUA D., PARAG A. PATHAK, AND CHRISTOPHER R. WALTERS (2013): "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 5, 1–27. [2525]

ANGRIST, JOSHUA D., PARAG A. PATHAK, AND ROMÁN ANDRÉS ZÁRATE (2019): "Choice and Consequence: Assessing Mismatch at Chicago Exam Schools," NBER Working Paper No. 26137. [2515,2529]

ANGRIST, JOSHUA D., ÒSCAR JORDÀ, AND GUIDO M. KUERSTEINER (2018): "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited," *Journal of Business & Economic Statistics*, 36, 371–387. [2533]

ARAI, YOICHI, YU-CHIN HSU, TORU KITAGAWA, ISMAEL MOURIFIÉ, AND YUANYUAN WAN (2022): "Testing Identifying Assumptions in Fuzzy Regression Discontinuity Designs," *Quantitative Economics*, 13, 1–28. [2515]

ASHENFELTER, ORLEY (1974): "The Effect of Manpower Training on Earnings: Preliminary Results," in *Proceedings of the 27th Annual Meeting of the Industrial Relations Research Association*, 252–260. [2510]

——— (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57. [2510]

ASHENFELTER, ORLEY, AND DAVID CARD (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648–660. [2510]

——— (Eds.) (1999): *The Handbook of Labor Economics*, Vol. 3A. Amsterdam: Elsevier. [2509]

BALKE, ALEXANDER, AND JUDEA PEARL (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. [2531]

BARNOW, BURT S. (1972): "Conditions for the Presence or Absence of a Bias in Treatment Effect: Some Statistical Models for Head Start Evaluation," Working Paper. [2510]

BARROW, LISA, LAUREN SARTAIN, AND MARISA DE LA TORRE (2020): "Increasing Access to Selective High Schools Through Place-Based Affirmative Action: Unintended Consequences," *American Economic Journal: Applied Economics*, 12, 135–163. [2515,2517]

BELLONI, ALEXANDRE, DANIEL CHEN, VICTOR CHERNOZHUKOV, AND CHRISTIAN HANSEN (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429. [2533]

BELLONI, ALEXANDRE, VICTOR CHERNOZHUKOV, AND CHRISTIAN HANSEN (2014): "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650. [2533]

BLOOM, HOWARD S. (1984): "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 8, 225–246. [2518]

BOUND, JOHN, DAVID A. JAEGER, AND REGINA M. BAKER (1995): "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450. [2533]

CALONICO, SEBASTIAN, MATIAS D. CATTANEO, MAX H. FARRELL, AND ROCÍO TITIUNIK (2017): "rdrobust: Software for Regression-Discontinuity Designs," *The Stata Journal*, 17, 372–404. [2533]

CARD, DAVID (2022): "Design-Based Research in Empirical Microeconomics," *American Economic Review*, 112, 1773–1781. [2533]

CATTANEO, MATIAS D., BRIGHAM R. FRANDSEN, AND ROCIO TITIUNIK (2015): "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate," *Journal of Causal Inference*, 3, 1–24. [2514]

CATTANEO, MATIAS D., MICHAEL JANSSON, AND WHITNEY K. NEWEY (2018): "Inference in Linear Regression Models With Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113, 1350–1361. [2533]

CATTANEO, MATIAS D., ROCIO TITIUNIK, AND GONZALO VAZQUEZ-BARE (2017): "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management*, 36, 643–681. [2514]

CHABRIER, JULIA, SARAH COHODES, AND PHILIP OREOPOULOS (2016): "What Can We Learn From Charter School Lotteries?" *Journal of Economic Perspectives*, 30, 57–84. [2525,2526]

CHAMBERLAIN, GARY (1984): "Panel Data," in *The Handbook of Econometrics*, Vol. 2, ed. by Zvi Griliches and Michael D. Intriligator. Amsterdam: Elsevier, 1247–1318. [2519]

CHERNOZHUKOV, VICTOR, DENIS CHETVERIKOV, MERT DEMIRER, ESTHER DUFLO, CHRISTIAN HANSEN, WHITNEY NEWEY, AND JAMES ROBINS (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [2533]

COHODES, SARAH R., ELIZABETH M. SETREN, AND CHRISTOPHER R. WALTERS (2021): "Can Successful Schools Replicate? Scaling Up Boston's Charter School Sector," *American Economic Journal: Economic Policy*, 13, 138–167. [2524]

COOK, THOMAS D. (2008): "Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics," *Journal of Econometrics*, 142, 636–654. [2510]

DALE, STACEY B., AND ALAN 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. [2517]

DAVIS, MATTHEW, AND BLAKE 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. [2529]

DE CHAISEMARTIN, CLEMENT, AND LUC BEHAGHEL (2020): "Estimating the Effect of Treatments Allocated by Randomized Waiting Lists," *Econometrica*, 88, 1453–1477. [2533]

DEHEJIA, RAJEEV H., AND SADEK WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [2533]

DOBBIE, WILL, AND ROLAND G. JR. 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. [2515]

DONG, YINGYING (2018): "Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs," *Oxford Bulletin of Economics and Statistics*, 80 (5), 1020–1027. [2515]

EPSTEIN, ISIDORE (1976): *Hebrew-English Translation of the Babylonian Talmud, Baba Bathra, Volume I*. London: Soncino Press. [2511]

FARR, WILLIAM (1853): *Weekly Return of Births and Deaths in London*. London: Registrar-General. [2509]

FRANDSEN, BRIGHAM R., LARS J. LEFGREN, AND EMILY C. LESLIE (2019): "Judging Judge Fixed Effects," NBER Working Paper No. 25528, February. [2531]

FROLICH, MARKUS, AND MARTIN HUBER (2019): "Including Covariates in the Regression Discontinuity Design," *Journal of Business and Economic Statistics*, 37, 736–748. [2514]

GALE, DAVID, AND LLOYD S. SHAPLEY (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–15. [2515]

GERARD, FRANÇOIS, MIIKKA ROKKANEN, AND CHRISTOPH ROTHE (2020): "Bounds on Treatment Effects in Regression Discontinuity Designs With a Manipulated Running Variable," *Quantitative Economics*, 11, 839–870. [2534]

GOLDBERGER, ARTHUR S. (1972): "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," Working Paper. [2510]

GRONAU, REUBEN (1977): "Leisure, Home Production, and Work—the Theory of the Allocation of Time Revisited," *Journal of Political Economy*, 85, 1099–1123. [2528]

HAHN, JINYONG (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [2533]

HAHN, JINYONG, PETRA TODD, AND WILBERT VAN DER KLAAUW (2001): "Identification and Estimation of Treatment Effects With a Regression-Discontinuity Design," *Econometrica*, 69, 201–209. [2512]

HANUSHEK, ERIC A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141–1177. [2510]

HANUSHEK, ERIK A. (1996): "School Resources and Student Performance," in *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, ed. by Gary Burtless. Washington, D.C.: Brookings Institution, 43–73. [2510]

HAUSMAN, JERRY A. (1983): "Specification and Estimation of Simultaneous Equation Models," in *The Handbook of Econometrics*, Vol. 1, ed. by Zvi Griliches and Michael D. Intriligator. North-Holland, 391–448. [2528]

HEARST, NORMAN, THOMAS B. NEWMAN, AND STEPHEN B. HULLEY (1986): “Delayed Effects of the Military Draft on Mortality,” *New England Journal of Medicine*, 314, 620–624. [2510,2519]

HECKMAN, JAMES J., AND EDWARD VYTLACIL (2005): “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669–738. [2531]

HIRANO, KEISUKE, GUIDO W. IMBENS, AND GEERT RIDDER (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161–1189. [2533]

HOXBY, CAROLINE M. (2000): “The Effects of Class Size on Student Achievement: New Evidence From Population Variation,” *Quarterly Journal of Economics*, 115, 1239–1285. [2512]

HUBER, MARTIN, AND GIOVANNI MELLACE (2015): “Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints,” *Review of Economics and Statistics*, 97, 398–411. [2531]

IDOUX, CLÉMENCE (2021): “Who Benefits From Selective School Attendance?” MIT Department of Economics, Working Paper, June. [2515]

IMBENS, GUIDO W., AND JOSHUA D. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475. [2518,2524]

IMBENS, GUIDO W., AND KARTHIK KALYANARAMAN (2012): “Optimal Bandwidth Choice for the Regression Discontinuity Estimator,” *Review of Economic Studies*, 79, 933–959. [2514,2516,2533]

IMBENS, GUIDO W., AND DONALD B. RUBIN (1997): “Estimating Outcome Distributions for Compliers in Instrumental Variables Models,” *Review of Economic Studies*, 64, 555–574. [2525,2531]

JOHNSON, STEVEN (2006): *The Ghost Map: The Story of London’s Most Terrifying Epidemic—and How It Changed Science, Cities, and the Modern World*. New York: Riverhead Books. [2510]

KING, JR. MARTIN LUTHER (1967): *Where Do We Go From Here: Chaos or Community?* Boston: Beacon Press. [2521]

KITAGAWA, TORU (2015): “A Test for Instrument Validity,” *Econometrica*, 83, 2043–2063. [2531]

KOLESÁR, MICHAL, AND CHRISTOPH ROTHE (2018): “Inference in Regression Discontinuity Designs With a Discrete Running Variable,” *American Economic Review*, 108, 2277–2304. [2533]

KRUEGER, ALAN B. (1999): “Experimental Estimates of Education Production Functions,” *Quarterly Journal of Economics*, 114, 497–532. [2510,2512]

LALONDE, ROBERT J. (1986): “Evaluating the Econometric Evaluations of Training Programs With Experimental Data,” *American Economic Review*, 76, 604–620. [2510]

LEAMER, EDWARD E. (1983): “Let’s Take the Con Out of Econometrics,” *American Economic Review*, 73, 31–43. [2534]

LEE, DAVID S. (2008): “Randomized Experiments From Non-Random Selection in US House Elections,” *Journal of Econometrics*, 142, 675–697. [2514]

LEE, DAVID S., AND THOMAS LEMIEUX (2010): “Regression Discontinuity Designs in Economics,” *Journal of Economic Literature*, 48, 281–355. [2510]

MOREIRA, MARCELO J. (2003): “A Conditional Likelihood Ratio Test for Structural Models,” *Econometrica*, 71, 1027–1048. [2533]

MOUNTJOY, JACK, AND BRENT HICKMAN (2020): “The Returns to College(s): Estimating Value-Added and Match Effects in Higher Education,” Becker Friedman Institute Working Paper. [2517]

NEWEY, WHITNEY K. (1985): “Generalized Method of Moments Specification Testing,” *Journal of Econometrics*, 29, 229–256. [2519]

NEWEY, WHITNEY K., AND KENNETH D. WEST (1987): “Hypothesis Testing With Efficient Method of Moments Estimation,” *International Economic Review*, 28, 777–787. [2519]

OLSON, LYNN, AND CRAIG JERALD (2020): “The Big Test: The Future of Statewide Standardized Assessments,” FutureEd, Georgetown University, April, https://www.future-ed.org/wp-content/uploads/2020/04/TheBigTest_Final.pdf. [2534]

ROBINS, JAMES M. (2000): “Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models,” in *Proceedings of the American Statistical Association*, Vol. 1999, 6–10. [2533]

ROSENBAUM, PAUL R., AND DONALD B. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55. [2533]

ROSENZWEIG, MARK R., AND KENNETH I. WOLPIN (1980): “Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment,” *Econometrica*, 48, 227–240. [2527]

RUBIN, DONALD B. (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701. [2517]

——— (1978): “Bayesian Inference for Causal Effects: The Role of Randomization,” *The Annals of Statistics*, 6, 34–58. [2517]

SIMS, DAVID (2008): “A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California,” *Journal of Policy Analysis and Management*, 27, 457–478. [2512]

SNOW, JOHN (1855): *On the Mode of Transmission of Cholera* (Second Ed.). London: Churchill. [2509]

STAIGER, DOUGLAS, AND JAMES H. STOCK (1997): "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586. [2533]

STOCK, JAMES H., AND MOTOHIRO YOGO (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by Donald W. K. Andrews and James H. Stock. Cambridge University Press, 80–108. [2533]

SUN, ZHENTING, AND KASPAR WÜTHRICH (2022): "Pairwise Valid Instruments,". arXiv:2203.08050. ArXiv preprint, March. [2531]

THERNSTROM, ABIGAIL, AND STEPHAN THERNSTROM (2004): *No Excuses: Closing the Racial Gap in Learning*. New York: Simon and Schuster. [2521]

THISTLETHWAITE, DONALD L., AND DONALD T. CAMPBELL (1960): "Regression-Discontinuity Analysis: An Alternative to the ex post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317. [2510]

WALTERS, CHRISTOPHER R. (2018): "The Demand for Effective Charter Schools," *Journal of Political Economy*, 126, 2179–2223. [2526]

WÜTHRICH, KASPAR, AND YING ZHU (2021): "Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis," *Review of Economics and Statistics*, 1–47. [2533]