Empirical Strategies in Economics: Illuminating the Path from Cause to Effect[1]

Prize Lecture 8 December 2021 by
Joshua D. Angrist, Massachusetts Institute of Technology, Cambridge, MA, USA

> 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*)

---

1. This is a revised version of my recorded Prize Lecture posted December 8, 2021. Many thanks to Jimmy Chin and Vendela Norman for their help preparing this lecture and to Noam Angrist, Hank Farber, Peter Ganong, Guido Imbens, and Parag Pathak for comments on an earlier draft. Thanks also go to my coauthors and Blueprint Labs colleagues, from whom I've learned so much over the years. 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 incorporates empirical findings from joint work with Atila Abdulkadiroğlu, Sue Dynarski, Bill Evans, Iván Fernández-Val, Tom Kane, Victor Lavy, Yusuke Narita, Parag Pathak, Chris Walters, and Román Zárate.

> 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 John Snow (1855) used to establish 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 in the battle to bring empirical strategies like Snow's into the econometric mainstream. Ashenfelter and Card's quest for an empirical strategy that reliably captures the causal effects of government training programs inspired me and others at Princeton to explore the econometrics of program evaluation.[2]

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 recognizes 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 differences between them after random assignment 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.[3]

This past October, I worried about what I should expect from the Economics Prize treatment effect. The spotlight and disruption

2. Their quest began in Ashenfelter (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 one Graduate Labor Economics class in 1986 by mentioning an intriguing study: Hearst et al. (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. From class, I went to the library, embarking on my first attempt to answer causal questions using observational data. Farr and Snow in the epigraph are quoted in Johnson (2006).

3. The 2019 Economics Prize awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer celebrates the rise of economics RCTs.

accompanying the prize made me wonder how the Economics Prize celebrity might change life for the Angrist family. It soon dawned on me that the matter of how public recognition affects a scholar's life is a simple causal question: the Economics Prize intervention is substantial, sudden, and well-measured; outcomes like health and wealth are easy to record. Although the Economics Prizes are probably not randomly assigned, a compelling empirical strategy for the Economics Prize treatment effect comes to mind, at least as a flight of empirical fancy.

Imagine a pool of Prize-eligible *applicants*: the group under consideration for the Prize. Applicants don't apply for the Prize themselves, they are nominated by peer scholars. You must be nominated to be awarded, and nominees are already a highly select group, so my fanciful Prize Impact Study looks only at nominee-applicants. This sample selection rule is but a first step. Credible applicants, I imagine, are evaluated by judges using criteria like publications, citations, nominating statements, and advisory letters of recommendation. I also imagine this material is aggregated and scored using some kind of rubric. With hundreds of applicants and much information used to score them, the scores are nearly continuous. Top scorers (up to 3 per field in any single year) are awarded a Prize.

Having identified the applicants and their scores, the next step in my Economics Prize Impact Study is to record the relevant *cutoffs*. The Economics Prize cutoff is the lowest score among those awarded a prize. Many the Economics Prize hopefuls just miss the cutoff. Looking only at near misses along with the winners, differences in scores between those above and below the cutoff begin to look serendipitous, almost randomly assigned. After all, near-Economics Prize's are among the most eminent of scholars too. With one more high-impact publication, or a little more support from nominators, they would have been awarded the Economics Prize gold. Some of them, someday, surely will be.

The empirical strategy sketched here employs a regression discontinuity (RD) design, one of applied econometrics' most powerful tools. RD exploits the jumps in human affairs induced by rules, regulations, and the need to classify people for assignment purposes. The Economics Prize turns on such a discontinuity. When a treatment or intervention is determined by whether a tie-breaking variable crosses a threshold, those just below the threshold become a natural control group for those who clear it. It now seems surprising to me that before the 1990s, economists had given this elegant idea little notice.[4]

4. The RD idea originated with psychologist Donald Campbell (Thistlethwaite and Campbell, 1960). Econometric pioneers Goldberger (1972) and Barnow (1972) discuss hypothetical applications of RD to evaluation of the nascent Head Start program. Cook (2008) and Lee and Lemieux (2010) summarize the intellectual history of RD.

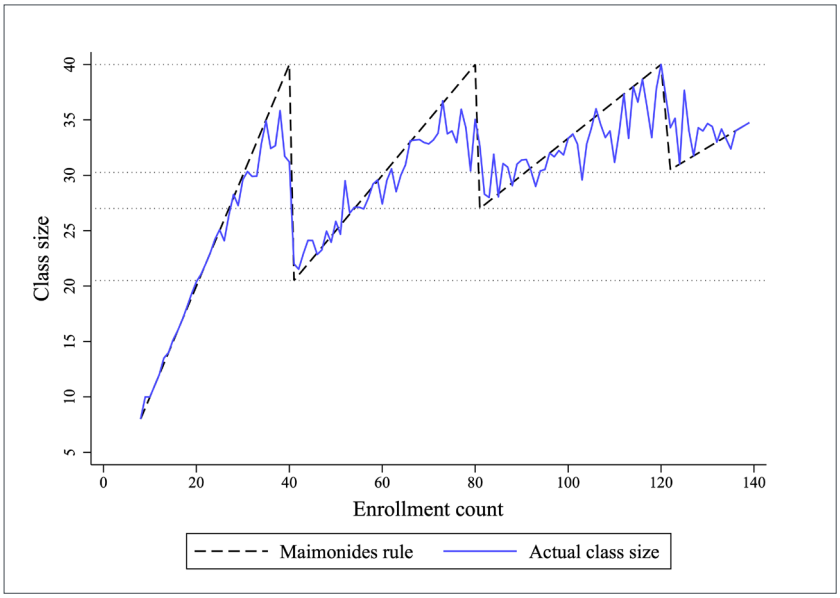


Figure 1. Class Size in 1991 by Initial Enrollment Count, Actual Average Size and as Predicted by the Maimonides Rule.

Notes: Average and predicted class size in Israeli 4th grade classes in 1991, conditional on enrollment. Predictions use Maimonides Rule.

RD does not require the 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. RD can allow, for example, for the fact that this year's near-Prize applicant might be next year's recipient. Allowing for this leads to the use of discontinuities in the rate at which treatment is assigned to construct instrumental variables (IV) estimates of the effect of treatment received. This sort of RD design is said to be *fuzzy*. But, as Steve Pischke and I wrote in our first book (in Angrist and Pischke (2009)), "fuzzy RD is IV."

The first RD application to which I contributed is Angrist and Lavy (1999), which exploits the rule used in Israeli elementary schools to determine class size. The causal effects of reduced grade-school class size have long preoccupied economists interested in the education production function. Much of this interest dates to the Hanushek (1986) survey of research on education production, which argues that education inputs like class size are at best weakly correlated with student achievement. Krueger's (1999) analysis of a rare class size RCT suggested that this finding may be an artifact of selection bias.

In the 1990s, Israeli classes were large. Students enrolled in a grade cohort of 40 were likely to be seated in a class of 40. But add another child to the cohort, making 41, and 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 Rambam proposed a maximum class size of 40.[5]

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. But the gist of the thing is a marked class size drop at each integer multiple of 40, the relevant cutoff, just 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.[6]

Lavy and I implemented the Maimonides Rule fuzzy RD research design in an IV set-up that can be described as follows. Writing f_j for the predicted 4th grade class size at school j, Rule-based enrollment is:

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

where r_j is the number of 4th graders at school j and $int\ (x)$ 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:

$$s_{ij} = \pi f_j + \rho_1 r_j + \delta_1' X_{ij} + \varepsilon_{ij}, (2)$$

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:

$$y_{ij} = \beta s_{ij} + \rho_2 r_j + \delta_1' X_{ij} + \eta_{ij}, (3)$$

Where β is the causal effect of interest and η_{ij} is the random part of potential achievement.

---

5. 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 proscribes 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)).

6. Or so they were in 1991 data. Revisiting the Maimonides Rule research design with Israeli data for 2002-11, Angrist et al. (2019a) estimates class size effects tightly distributed around zero. Many countries have their own version of the Maimonides Rule, usually with cutoffs below 40. For example, Angrist et al. (2017a) uses Italy's version to estimate causal effects of class size on the manipulation of standardized test scores. Sims (2008) uses the Maimonides Rule to document unintended consequences of a California class size reduction program: the program encouraged the use of "combination classes" mixing elementary school grades where this led to a reduction in average class size.

Angrist and Lavy (1999) uses the local average treatment effects (LATE) framework to interpret IV estimates based on (2) and (3) in a world of heterogeneous potential outcomes. Following a suggestion from Caroline Hoxby, we also undertook an analysis of applicants in "discontinuity samples" limited to applicants close to Maimonides Rule cutoffs.[7] Around the same time, Hahn et al. (2001) formalized the LATE interpretation of nonparametric fuzzy RD. Applications of this new approach to IV and RD, initially isolated, bloom widely today.

This lecture uses examples to illustrate the power of IV and RD empirical strategies to uncover new causal knowledge. Most of my examples concern the achievement effects of attendance at schools of various kinds. The question of school effects highlights key features of the LATE framework, including an extension to distribution treatment effects. This extension shows how urban charter school attendance closes Black-white achievement gaps. The last example supports a surprising exclusion restriction: diversion from high- performing urban charters explains why enrollment at Chicago's selective enrollment high schools reduces student achievement. The lecture concludes with a few comments on the evolution of empirical economics.

2. EXAM TIME!

Would a comparison of Economics Prize laureates to near-laureates *really* be a good natural experiment? This claim seems more compelling for comparisons of schools with 40 and 41 fourth graders than for near- and officially-recognized laureates. Yet, both scenarios exploit a feature of the physical world: provided the tie-breaking variable (known to RD mavens as the "running variable") has a continuous distribution, assignment rates approach 0.5 when computed in a narrow window around the cutoff used to adjudicate awards. In RD empirical work, the window around such cutoffs is known as a *bandwidth*. Importantly, the limiting win rate is 0.5 for everybody, regardless of how qualified they look going into the Economic Prize competition.

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.[8] Only applicants ranked highly enough are offered a screened-school seat. In other words, the admissions regime for screened schools is a lot like the scheme I've imagined for the Economics Prize.

---

7. In work concurrent with ours, Hoxby (2000) uses population variation to construct instruments for class size.

8. More precisely, this share refers to school programs—school buildings may host more than one program.

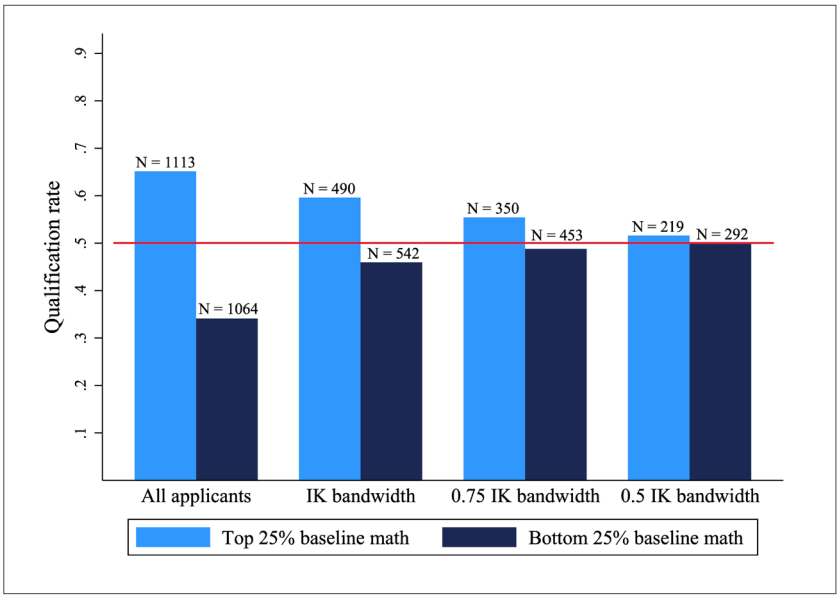


Figure 2. Qualification Rates Near the Townsend Harris Cutoff.

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.

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. The first pair of bars in Figure 2 show qualification rates *conditional* on a measure of pre-application "baseline" achievement. 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 there than are applicants with low baseline scores. In a shrinking symmetric 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 .75 IK and then .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 tossing a coin rather than by selecting only those who scored highly on the school's entrance exam. Yet, even when admissions operate by the latter rule, the data can be arranged so as to mimic the former.[9]

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 Laureates between them.[10] Exam-school proponents see these schools as democratizing public education. Wealthy families, they argue, can access exam-school 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 collaborators and I have examined the causal effects of exam

9. The figure illustrates the following theorem. Suppose applicant i qualifies when running variable Ri clears a fixed cutoff, gr. tau, and that the distribution of Ri is continuously differentiable. Let Qi = 1[Ri > gr. tau] indicate qualification and let Wi be a random variable (like baseline scores) unchanged by qualification. Then,

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

Qualification is but one input determining the conditional probability of assignment in New York's city-wide high school match. Abdulkadiroğlu et al. (2017a, 2022) derive the distribution of school assignments generated by the NYC high school match. As far as I know, Cattaneo et al. (2015) is the first empirical application of the local random assignment interpretation of RD. See also Frolich and Huber (2019) and Cattaneo et al. (2017).

10. Townsend Harris has graduated three Laureates, including economist Kenneth Arrow.

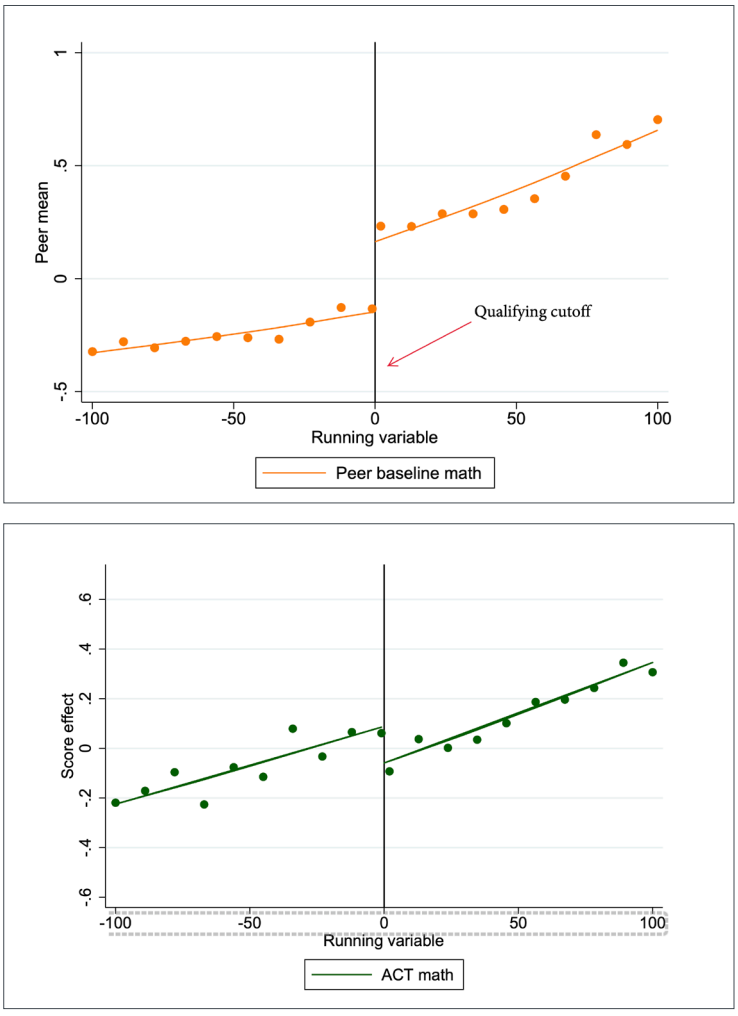


Figure 3. Peer Baseline and ACT Math Effects at Qualifying Cutoffs for Chicago Exam Schools.

Notes: This figure plots peer baseline math scores (Panel A) and ACT math scores (Panel B) against the exam school admissions composite. 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 among the Chicago Public School district's test-taking population. A student's peers are all the other 9th graders enrolled at the same school. The running variable is centered around the qualifying cutoff. Applicants who clear their 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 (LLR). The LLR uses a triangular kernel and the kernel bandwidth is computed as suggested by Imbens and Kalyanaraman (2012). All variables are plotted after partialing out saturated qualifying-cutoff-by-tier-by-application-year fixed effects.

school attendance in Boston, Chicago, and New York.[11] This work has generated surprising 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 et al., 2014). This refers to the fact that, while exam school students undoubtedly have high test scores and other good outcomes, this is not a causal effect 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 good performance of exam school students reflects *selection bias,* that is, 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 et al. (2017b) and Angrist et al. (2019b)).[12] The left panel of Figure 3 explains why exam schools are so attractive to parents. This panel plots *peer mean achievement* – that is, the 8th grade test scores of an applicants' 9th grade classmates–against the admissions tie-breaker, for a subset of applicants to any one of Chicago's nine 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 any 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 Prize awarded to Alvin Roth and Lloyd Shapley). As it happens, the Chicago DA implementation is well-approximated by 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*.[13]

11. David Autor, Parag Pathak, and I founded Blueprint Labs in 2011 (originally, the MIT School Effectiveness and Inequality Initiative). Since then, lab staff and research assistants have provided an incomparable framework for research on education and the labor market. Time flies when you're having fun!

12. Other Blueprint exam-school research includes Angrist and Rokkanen (2015), Idoux (2021), and Abdulkadiroğlu *et al.* (2022). Dobbie and Fryer (2014) and Barrow *et al.* (2020) also use RD to study exam schools in New York and Chicago, respectively.

13. 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 rank 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.

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 reflects 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 make it 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 (on tests 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.[14] 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

The LATE framework offered a new understanding of the results of empirical strategies involving IV and RD. The prize that Guido Imbens and I share is in recognition of the growing importance of this conceptual framework. In his latest book, cognitive psychologist Steven Pinker (2021) writes: "When a data scientist finds a regression discontinuity or an instrumental variable, it's a really good day." I like to think we made such days even better.

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, draft lottery numbers randomly assigned in the 1970s, generates variation in Vietnam-era veteran status (Angrist, 1990). The second, 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).

---

14. Barrow et al. (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.

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 younger, and are therefore allowed to drop out of high school (on their 16th birthday) with less schooling completed than those born later. But 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 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 a 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 b be the effect of the 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. Yet, to this day,

I'm often asked how it can be true that in a scenario where treated subjects selectively decline treatment, the average causal effect on the treated is knowable. Remarkably, Bloom derived this result from first principles, with no connection to IV.

The LATE theorem (Imbens and Angrist (1994) and Angrist et al. (1996)) generalizes the Bloom theorem. 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.[15] 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 inequality doesn't matter). Given this restriction, which we called *monotonicity*, LATE says:

$$E[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}] = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \quad (4)$$

$$= \frac{\delta}{\pi_1 - \pi_0}, \quad (5)$$

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 (4) 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; Angrist et al. (2000) covers continuous treatments and simultaneous equations models.

Ice Cream at Princeton, AIRtime at Harvard

At Princeton and Harvard, I read and reread Chamberlain (1984), Newey (1985), and Newey and West (1987). These articles answered most of my questions about econometric theory as it relates to estimation and inference. But not all. So, I was lucky to be able to call on Whitney Newey (a Ph.D. advisor) and Gary Chamberlain (a colleague) 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. In 1990, I apprenticed to Gary as co-instructor in his econometrics class, an experience from which I learned at least as much as our students.

Angrist (1990) uses draft-lottery dummies as instruments for veteran status in a two- sample linear IV procedure detailed in Angrist and Krue-

15. We owe this elegant 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 to model latent treatment assignments. Gary suggested we define D_{0i} and D_{1i} without recourse to a latent-index model.

ger (1992). Motivated by the fact that Hearst et al. (1986) looked at veteran effects on mortality, a dummy dependent variable, I also sought an empirical strategy for IV in a qualitative response model. With little beyond bivariate probit to show for my efforts, Newey suggested I seek new causal knowledge from biostatistics maven Jamie Robins at the Harvard School of Public Health.[16] 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 as well. It was Rubin who put us "on AIR," in Angrist et al. (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 that instrumental variables are a formidable tool in the quest for new causal knowledge.

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 much longer letter that marked the beginning of our 3-way collaboration. He agreed that the draft lottery 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 Don] 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." This double-indexing of potential outcomes allowed us to separate exclusion restrictions from independence assumptions, a feature of the LATE framework adopted in Angrist et al. (1996).

16. Angrist (1991) shows that, when the linear first stage implied by a latent-index model is a non-zero constant, the IV estimand for a single dummy instrument is an 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.

3.1 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 framework here through a research question that has occupied me for almost two decades: the causal effect of charter school attendance on learning.[17]

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 awarded 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 regulate pay and working conditions, often in a very detailed manner.

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, with over 80% of KIPP students poor enough to qualify for the federal government's subsidized lunch program.[18]

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

17. My interest in charter schools dates to 2003, when Michael Goldstein, then CEO of the Match Charter High School, 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 couldn't get the requisite data on test scores. The first charter lottery analysis to which I contributed was released in 2009 (Abdulkadiroğlu et al., 2009), later published as Abdulkadiroğlu et al. (2011).

18. 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." Quantitative analysis of the No Excuses paradigm begins with Thernstrom and Thernstrom (2004).

reflects the fact that KIPP attracts families whose children are more likely to succeed anyway.

A randomized trial might prove decisive in the debate over attendance effects at schools like KIPP. Alas, like the Economics prizes, seats at KIPP are not randomly assigned. At least, not entirely. In fact, Massachusetts charter schools with more applicants than seats offer their seats by lottery. Specifically, applicants are ordered randomly, and charter school seats filled by making offers down this randomly ordered list. Some of these offers are ignored, while some students way down the wait list nevertheless find their way to a seat at KIPP. By and large, however, the *opportunity* to attend KIPP is randomly assigned.

A decade ago my collaborators and I collected data on KIPP Lynn middle school admissions lotteries, laying the foundation for charter school research published in Angrist et al. (2010a, 2012). At the time, the KIPP middle school in Lynn, Massachusetts was the first of its kind in New England. Some KIPP applicants bypass the lottery—those with previously enrolled siblings are guaranteed admission, while a few applicants 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 failed to enroll come September. Some had moved away, while others ultimately preferred a traditional public school. Among those 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 at KIPP nevertheless found their way into KIPP. The effect of an offer on KIPP enrollment rates is $199/253 - 5/118 \approx 0.74$. 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 analysis sketched here looks at KIPP attendance effects on test scores for tests taken at the end of the grade following the application grade (i.e., 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, data on scores have been 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 (written $-.3\sigma$).

Among participants in KIPP entry lotteries, applicants offered a seat had standardized math scores close to zero (-0.003 to be precise), that is,

near the state mean. Because KIPP applicants start with 4th grade scores that average roughly .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 much more typical result for 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, so we can be confident this isn't a chance finding. What does an *offer* effect around .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 equals one for KIPP applicants who receive offers and zero otherwise. As in the 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 variable we care about, 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 the omitted variables we might like to control for, in this case, variables like KIPP applicants' family background or motivation to enroll. This is called the *independence assumption.*
III. Finally, IV logic requires an *exclusion restriction.* The exclusion restriction postulates a single measured channel through which the instrument, Z_i, affects outcomes. Here, the exclusion restriction amounts to the claim that the 0.355σ score differential between lottery winners and losers is entirely attributable to the .74 win-loss difference in attendance rates, that is, to the effect of Z_i on D_i.

IV empirical strategies characterize 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:

Effect of offers on scores =
{Effect of offers on attendance} × {Effect of attendance on scores}.

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 restatement of equation (5), expressed here in words.

The effect of the instrument (offers) on outcomes (scores) plays a central role in the IV story and therefore has a special name: this is the *reduced form*, denoted by d in (5). Dividing the reduced form (0.355σ) by the first stage, the KIPP attendance effect works out to be:

$$.48\sigma \approx \frac{.355\sigma}{.745} = \frac{(-0.003\sigma) - (-0.358\sigma)}{0.787 - 0.042}. \quad (7)$$

Almost half a standard deviation gain in math scores is a remarkable result. Few education- related interventions have such large effects.[19]

It's one thing to be able to compute an IV estimate and another to know what it means. Children 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 Lynn or a Lynn 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.

		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*)

Table 1. The Four Types of Children.
Notes: KIPP = Knowledge is Power Program.

19. 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 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.

As I've mentioned, LATE theory is illuminated by the biblical story of Passover, which explains that there are four types of children, each with characteristic behaviors. Table 1 classifies applicants named Alvaro, Normando, and Camila, as 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 1 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 only 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_{1i}. Potential choices are like potential outcomes: for any given applicant, we see only one or the other.

Never-takers and always-takers like Normando and Alvaro 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 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, but stoically accept the verdict if they lose. Camila complies with her lottery offer, attending KIPP when she wins but not otherwise. In other words, Camila has $D_{1i} = 1$, $D_{0i} = 0$.

The term complier links IV with the RCTs we hope to mimic. Many randomized trials randomize only the *opportunity* to be treated, while the decision to comply with the treatment protocol remains 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 always-takers like Alvaro and never-takers like Normando because the instrument is unrelated to their treatment status.

The defiers in Table 1 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. Note the distinct roles each IV assumption plays 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 assert 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 components lead to a simple formula for causal effects on compliers.

The LATE framework is sometimes seen as limiting the relevance of econometric inference. Yet, the population of compliers is 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 legislated charter expansion is central to the education policy debate (since the founding of Blueprint Labs, we've had two ballot initiatives on this matter). Cohodes et al. (2021) tackles the question of whether IV 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 be a remarkably reliable guide to performance of newly-opened campuses.

No Excuses for Not Closing the Achievement Gap

The LATE framework, meaning the assumptions behind the theorem, identifies the entire distribution of potential outcomes for compliers. To see this, suppose first that treatment, D_i, is randomly assigned in a stratified randomized trial, with strata encoded by X_i. Conditional random assignment implies that:

Skriv en ekvation här.

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i | X_i. \quad (8)$$

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

$$\begin{aligned} E[Y_i|D_i = 1, X_i] - E[Y_i|\text{Di} = 0, \text{Xi}] &= E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 0, X_i] \quad (9) \\ &= E[\text{Y1i}|X_i] - E[Y_{0i}|X_i] \\ &= E[Y_{1i} - Y_{0i}|X_i]. \end{aligned}$$

Because the logic of (9) works for any function of Y_i, we can replace Y_i with Y?i?*? (c) for any constant, c. This substitution yields:

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

The right-hand side of this expression is the difference in the distributions of Y_{1i} and Y_{0i} within strata, evaluated at c. Thus, RCTs reveal the entire distribution of each potential outcome, as well as the difference in potential-outcome distributions at any point. Such distributional comparisons are seen often in evaluations of life-saving vaccines and treatment regimens, where the distribution of interest is that of survival time.

The LATE analog of the conditional independence expressed by (8) is the statement that

$$\{Y_{1i}, Y_{0i}\} \; D_i | Y_{0i} > D_{0i}. \qquad (10)$$

This holds because, by virtue of monotonicity, $Z_i = D_i$ for compliers, so

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

Independence and exclusion imply that the right-hand side here is just the marginal mean of Z_i. Conditional independence relation (10) is remarkable because, unlike in an RCT, D_i itself is not taken to be randomly assigned in the LATE framework. Yet, for compliers, D_i is independent of potential outcomes and therefore as good as randomly assigned. This suggests we can learn all we might like to know about the distributions of potential outcomes for compliers. Of course, 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).[20]

Although the theory behind this is necessarily technical, the value of a LATE analysis of distributions is easily appreciated in practice. Recall that the KIPP study summarized above is motivated in part by Black-white achievement gaps. The top of Figure 4 presents some context for this concern by depicting the distribution of 4th grade scores for four cohorts of applicants to Boston charter middle schools. The two panels in the upper part of the figure show score distributions by race, tabulated separately for treated and untreated compliers. Treated compliers are compliers who attended a charter school, while untreated compliers did

20. The cumulative distribution function of Y_{ji}; $j = 0, 1$ can be consistently estimated using:

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

where, as before $Y_i^*(c) \equiv 1(Y_i < c)$ Densities can be obtained by replacing indicator functions with kernels; see Angrist *et al.* (2016) for details.

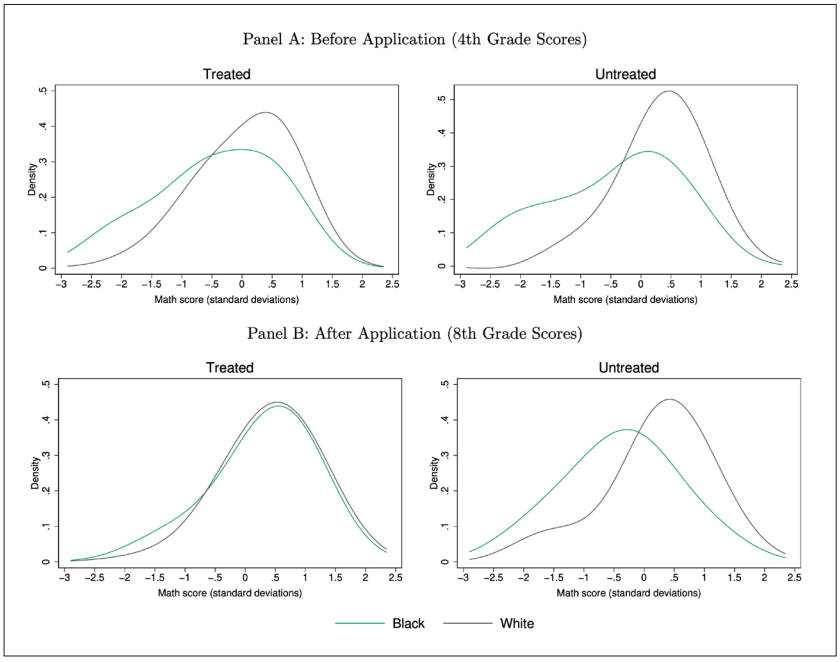


Figure 4. Charter Schools Close the Achievement Gap

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 *et al.* (2018).

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 8th grade, the distribution picture has changed markedly. This can be seen at the bottom of Figure 4, which again shows score distributions by race, separately for treated and untreated compliers. In eighth grade, treated compliers have completed middle school at a Boston charter. Like KIPP Lynn, these schools mostly adhere to a version of 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 enrollment. Consequently, Black charter students, who start middle school with lower baseline scores, see their learning accelerated more by charter attendance than do whites.[21] This differential impact is reflected in

21. For alternative views of this pattern, see Angrist et al. (2013) and Chabrier et al. (2016).

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 the score distributions for untreated compliers, by contrast (shown at the bottom right of the figure), have changed little from baseline, with whites still clearly ahead of Blacks.

3.2 Where Do Babies Come From?

Economists have long been interested in the contribution of childbearing to gender gaps in earnings and hours worked. Bill Evans and I used two IV empirical strategies to capture causal effects of childbearing on parents' labor supply. The LATE framework implies that these two instruments, though applied to the same causal relationship, need not identify the same average causal effect. The population of compliers is instrument-specific and different sorts of compliers are 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, at least in part, on where the babies in question 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 be more likely to have children than mothers with strong labor force attachment or high earnings potential. And the causality might just as well run the other way, with increased female employment driving down fertility. This makes the observed association between family size and employment hard to interpret.

Angrist and Evans (1998) solves these omitted-variable and endogeneity problems using instrumental variables that affect the birth of a third child. Our first 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 twinning 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 the sense that, when first and second-born children are both boys or both girls, the likelihood of a third child jumps. Angrist and Evans (1998) deploys the twins and same-sex empirical strategies in samples from the 1980 and 1990 US Census public use files.

			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: $\chi2(1)$	—	—	—	—	—	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: $\chi2(1)$	—	—	—	—	—	3.5
	(p-value)						(0.06)

Table 2. IV Estimates of the Effects of Family Size on Labor Supply.
Notes: The 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) extract from the 1980 U.S. census 5 percent sample, including 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.

The twins first stage in 1980 Census data is about .6, an estimate reported in column 2 of Table 2 (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 essentially random, unrelated to potential outcomes or demographic characteristics, and that a multiple birth affects labor supply solely by increasing fertility[22] 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 pair. This is documented in column 4 of Table 2 (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 labor supply solely by increasing fertility.

IV estimation of third-birth effects using the twins instrument yields a precisely-estimated reduction in weeks worked of a little over 3 weeks, with an employment reduction of about .08 points. These results, which appear in column 3 of Table 2, are smaller in absolute value than the corresponding ordinary least squares (OLS) estimates reported in the first column. The latter, computing using a set of controls listed in the table note, show roughly 9 fewer weeks worked and around an 18 percentage point reduction in employment for mothers who have a third birth. In view of the IV estimates, these large OLS estimates seem likely to be exaggerated by selection bias.

IV estimates constructed using the same-sex instrument, reported in

22. These conditions are unlikely to be met in a contemporary sample because the twin birth rate is boosted by in-vitro fertilization and related fertility treatments. Fertility interventions like these are more common among older and more educated women. IVF rose to prominence in the mid-1990s.

column 5 of Table 2, are substantially more negative than the corresponding twins IV estimates, though still smaller in magnitude 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 2 reports two-stage least squares (2SLS) estimates of childbearing 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-at-a-time IV estimates capture the same causal effect. This test generates p-values of .02 and .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.[23]

In Angrist and Fernández-Val (2013), my former Ph.D. student Iván Fernández-Val and I argue that smaller twins-IV estimates of fertility effects reflect differences between 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. By LATE logic, therefore, 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,

$$\begin{aligned} E[Y_{1i} - Y_{0i}|D_{1i} = 1, D_{0i} = 0] &= E[Y_{1i} - Y_{0i}|D_{0i} = 0] \quad (11) \\ &= E[Y_{1i} - Y_{0i}|D_{0i} = 0, Z_i = 0] \\ &= E[Y_{1i} - Y_{0i}|D_i = 0]. \end{aligned}$$

In other words, twins instruments reveal the effect of a third birth on women who choose to have small families (the second equals sign above uses the independence assumption; the third uses the fact that women who either have twins or for whom D0i = 1 have a third child). The same-sex instrument, by contrast, captures childbearing effects on women who can be nudged into additional childbearing by the sex mix of their first- and second-born.

Differences between same-sex and twins compliers are economically important because women who choose smaller families are especially likely to be college-educated. College education and the consequent higher pay this brings encourages educated mothers to use paid childcare. This in turn facilitates labor force participation in the wake of a third birth. Same-sex compliers, by contrast, are only about two-thirds as likely as the typical mother of two to have a college degree and are therefore

23. 2SLS combines multiple instruments by using the fitted values from the first-stage equation with all instruments on the right hand side as a single combined instrument. Most IV estimates are computed using 2SLS because 2SLS neatly accommodates covariates, while also combining multiple instruments efficiently. Models using 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, e.g., Hausman (1983) for details.

less likely than twins compliers to avail themselves of extra paid childcare in response to a third birth. The work-reducing consequences of childbirth for women providing home care is higher than for women using paid care.[24]

This tale of two instruments shows how LATE can be used to reconcile disparate results from two natural experiments, even while both experiments identify features of the same underlying causal relationship. And, having seen how the relevant compliant populations differ, economic reasoning suggests a theoretically-grounded explanation for why these differences in maternal characteristics lead to differences in impact.

4. CONSTRUCTING CAUSAL STORIES

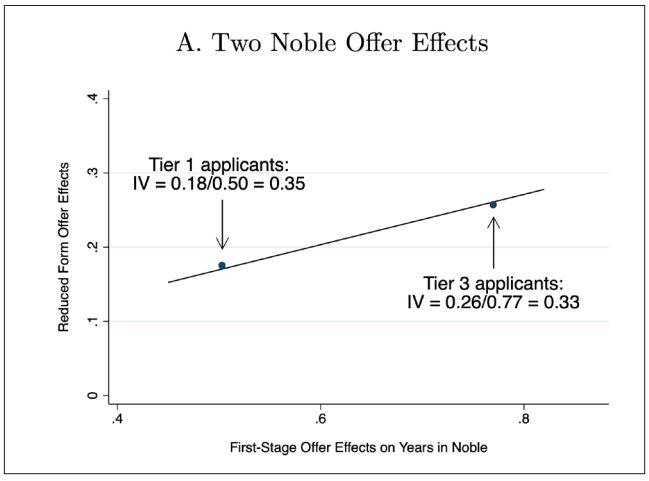
My Blueprint Labs colleagues and I have uncovered many surprising causal stories. I'll finish this lecture with one of the most intriguing, a story that resolves the puzzle of negative Chicago exam school effects. As a reminder, the challenge here is to explain why enrollment at one of the Windy City's coveted selective enrollment high schools appears to reduce learning rather than increase it.

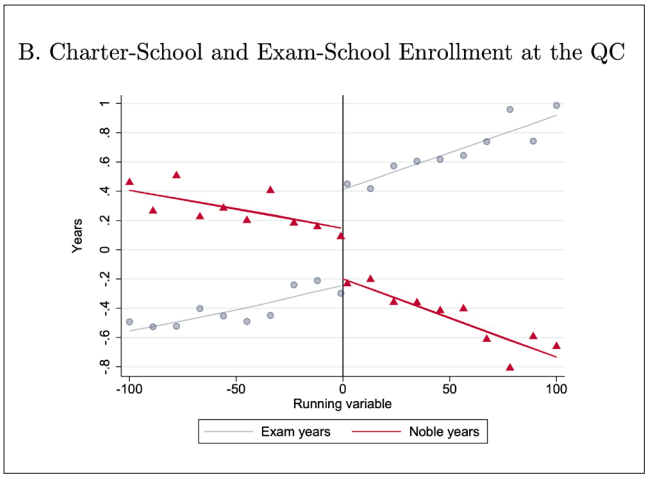
Economic reasoning is all about alternatives. What's the alternative to an exam school education? For most applicants to Chicago exam schools, the leading non-exam-school alternative is a traditional public school. But many of Chicago's rejected exam school applicants enroll in charter schools. Exam school offers therefore reduce the likelihood of charter-school attendance. Specifically, exam-school offers divert applicants away from high schools in the Noble Network of charter schools. Noble, with pedagogy much like KIPP's, is one of Chicago's most visible charter providers, enrolling 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 offered a Noble seat in a lottery naturally spend more time enrolled there than applicants not offered a seat. They also have higher ACT scores as a result (Davis and Heller (2019) is the first study using lotteries to document Noble effectiveness).

---

24. College graduation rates among compliers can be compared using the fact that the probability a complier has Bernoulli characteristic xi = 1, relative to the marginal probability that xi = 1, is given by the ratio of the first stage conditional on xi = 1 to the unconditional first stage. See Angrist and Pischke (2009) for details. A second feature separating the two complier populations is the difference in average age of the second-born. Multiple births produce second- and third-born children of the same age, while same-sex compliers can space births at leisure. Twins compliers therefore have exceptionally young second-borns at the time of a third birth. This moderates the work-reducing consequences of a third birth. The economic reasoning behind this argument draws on Gronau (1977).





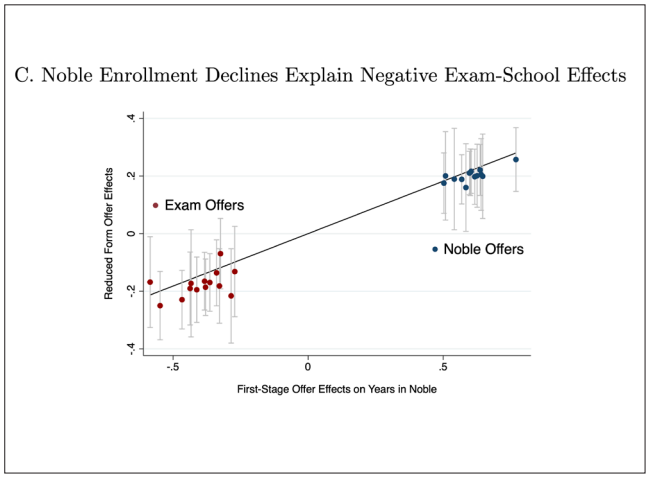


Figure 5. Explaining Chicago Exam School Effects with Charter Enrollment.

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. Panel C is a visual instrumental variable (VIV) plot of exam and Noble offer effects for a set of 14 covariate-defined groups. Exam effects are in red; Noble effects are in blue. Covariate-specific estimates are computed one at a time in the relevant subsamples. The slope of the line through these estimates is 0.34 in Panel A and 0.36 in Panel C. Fitted lines are forced to pass through the origin. Whiskers in Panel C mark 95% confidence intervals.

This pattern is captured in Panel A of Figure 5.[25] The x-axis in Panel A shows lottery effects of lottery offers on years enrolled at Noble; this is the Noble first stage for an IV setup that uses a dummy indicating Noble lottery offers as an instrument for Noble enrollment. I switch here to years enrolled rather than a dummy indicating any 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 first stage effects for two groups, one for Noble applicants who live in one of Chicago's lowest-income neighborhoods (labeled "Tier 1") and one for Noble applicants who live in a higher-income area ("Tier 3").[26]

Recall the IV chain reaction: the reduced-form effect of an instrument (in this case, a Noble offer) on outcomes (here, ACT math scores) equals the causal effect of interest (Noble enrollment) 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, also labeled in the figure. In this case:

$$Effect\ of\ Noble\ enrollment\ on\ ACT\ scores = \frac{\{Effect\ of\ Noble\ offers\ on\ ACT\ scores\}}{\{Effect\ of\ Noble\ offers\ on\ Noble\ enrollment\}}$$

For Tier 1, the relevant numbers are $0.35 = \frac{0.18}{0.50}$, while for Tier 3 we have $0.33 = \frac{0.26}{0.77}$.

For Noble applicants from both tiers, these first-stage and reduced-form effects imply an impressive yearly Noble enrollment impact of about a third of a standard deviation. Importantly, the line drawn through these two points runs through the origin (though the origin isn't included in the plot area covered by figure). Because 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 estimates, at 0.34s). This is the same as saying that reduced-form effects are proportional to first-stage effects, with the same factor of proportionality. Why is this important? By fitting the origin as well as the two points in the figure, we've substantiated an exclusion restriction which says that, given an applicant group for which Noble offers are unrelated to Noble enrollment, we should expect to see no reduced-form effect of these offers on test scores.

How *consistent* is the evidence for a Noble enrollment effect on the order of .34s per year? The blue points plotted in the upper right area of

25. Like Figure 3, this is derived from Angrist et al. (2019b).
26. Most Chicago public school students are low-income and non-white; tiers classify relative income within the city population.

Panel C of Figure 5 show first-stage and reduced-form Noble offer effects for 12 additional groups (2 more tiers and 12 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 0.36s, close to the slope of the line in Panel A. Again, the fitted line passes through the origin.

The fact that the line fits reasonably well bears a digression. As noted in the discussion of twins and same-sex instruments, over-identification tests compare alternative IV estimates of the same causal effect. In a constant-effects framework, alternative IV estimates of the Noble enrollment effect should be similar unless one of the instruments is invalid. Yet, as we've seen, LATEs using different instruments can differ even when all instruments are valid. Even in the LATE framework, however, reduced-form effects are scaled by the size of the corresponding first stage. 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 likewise be zero. These restrictions hold even in the absence of constant causal effects.[27]

What do the Noble IV estimates in Panel A of Figure 5 have to do with the effects of *exam-school* enrollment? The answer appears in Panel B, which complements the RD plots in Figure 3 with an added twist. The gray line in Panel B shows, as we should expect, that exam school enrollment jumps for applicants who clear their qualifying cutoff (qualification implies an exam-school applicant is offered an exam-school seat somewhere). The effect of qualification on any exam-school enrollment is about 21 percentage points, and about 0.61 for years enrolled. At the same time, in this sample of applicants who also applied to a Noble campus, Noble enrollment falls by about 15 percentage points at the qualifying cutoff (years enrolled at Noble falls by 0.37). This is the diversion effect of exam school offers on Noble enrollment.

IV affords us the opportunity 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 effect that exam-school offers have on *Noble* enrollment. In this account of Chicago exam school effects, exam school offers leave achievement (and other outcomes discussed in Angrist et al. (2019b)) unchanged for those not diverted from Noble.

In support of this claim, note first that the points plotted in red in

27. 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 et al. (2019). Angrist et al. (2010b) uses the "no first stage, no reduced form" restriction to assess the validity of instruments for family size.

Panel C of Figure 5 are all 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 a particular group of applicants. Because exam-school offers divert many exam-school applicants away from Noble, these numbers are negative (as with the blue points, there's a red point for each of the 14 groups defined by tier and demographic characteristics).

We've already seen that Noble applicants offered a Noble seat realize large ACT math gains as a result. Now consider exam-school offers as an instrument for Noble enrollment. As always, IV is a chain reaction. 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 in Panel C of Figure 5, we should expect reduced-form effects of exam school qualification equal to about −.13s. This is roughly consistent with the 14 reduced-form estimates plotted in red at the bottom left of Panel C (the fit isn't perfect; the reduced-form qualification effects in the figure cluster closer to −0.15 than to −0.13, but this is well within the range implied by sampling variance).

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 instrumental variable. Importantly, as in Panel A, the line plotted in Panel C runs through the origin. The IV story leaves us totally committed: in applicant groups where exam-school offers have little or no effect on charter-school enrollment, these offers should leave ACT scores unchanged.

5. EMPIRICAL ECONOMICS GETS SERIOUS

I computed the IV estimates in my Princeton Ph.D. thesis on a mainframe monster using 9-track tapes and leased spaced on a communal hard drive. Princeton graduate students mastered IBM job control language, the better to manipulate tape reels the size of a cheesecake (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 economics' shift towards transparent empirical strategies applied to concrete causal questions, like the questions my co-laureate David Card has examined so convincingly. The shift towards question-driven rather than model-driven empirical work fueled a wave of econometric innovation that continues today.

Much of the question-driven methodological agenda builds on Rosenbaum and Rubin's (1983) propensity score theorem. This theorem changed applied econometrics by focusing our attention on the process

determining treatment assignment rather than on models for outcomes. Dehejia and Wahba (1999) was the first to demonstrate the value of this approach, while Hahn (1998) and Hirano et al. (2003) raised new theoretical questions about the score. More recently, Belloni et al. (2014) use machine learning to model the score while also modeling outcomes. This work can be seen as extending the Robins (2000) notion of double robustness to a wider class of empirical strategies.

The flowering of distinctive RD methodology is ongoing. In a cascade of contributions, econometricians continue to tackle the vexing details of nonparametric RD bandwidth choice (as in Imbens and Kalyanaraman (2012) and Calonico et al. (2017)). Nonparametric RD also requires a modicum of continuity – this fact might handicap the fanciful Prize treatment effects study that I began with. Yet, Kolesár and Rothe (2018) show we can make do with a discrete running variable. De Chaisemartin and Behaghel (2020) solve estimation problems arising in RD designs when cutoffs are behaviorally determined, 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 on the finite-sample behavior of IV estimators. Progress here is summarized in Andrews et al. (2019). In Angrist and Kolesár (2021), Michal Kolesár and I argue that, when it comes to just-identified IV, at least, worries about bias are overblown.

I'm looking forward to solutions to the many problems my labmates and I encounter in our empirical work on causal effects. These include development of a complete estimation and inference framework for studies combining research design with market design (Abdulkadiroğlu et al., 2017a, 2022). Inference with clustered data remains as vexing as ever, though Abadie et al. (2017) makes the clustering question easier to address. RD is not foolproof: working on Angrist et al. (2019a), I was surprised to learn that school enrollment is an easily-manipulated running variable. More and better solutions for this problem, as in Gerard et al. (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 yet 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 elite illusion in school ratings (Angrist et al., 2017b, 2021a,b). Yet, without assessments of reading skills, how are we to know whether schools are teaching children 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 grad school 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 tell 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, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association,* **97**, 284–292.

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

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When Should You Adjust Standard Errors for Clustering?" NBER Working Paper No. 24003, November.

Abdulkadiroğlu, A., J. D. Angrist, S. R. Cohodes, S. M. Dynarski, J. Fullerton, T. Kane, and P. A. Pathak (2009): "Informing the Debate: Comparing Boston's Charter, Pilot and Traditional Schools," *The Boston Foundation.*

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

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

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

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

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

Abdulkadiroğlu, A., J. D. Angrist, Y. Narita, P. A. Pathak, and R. A. Za´rate (2017b): "Regression Discontinuity in Serial Dictatorship: Achievement Effects at Chicago's Exam Schools," *American Economic Review: Papers & Proceedings*, **107**, 240–245.

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

Angrist, J. and M. Kolesa'r (2021): "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV," NBER Working Paper No. 29417, October.

Angrist, J. D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review,* **80**, 313–336.
— (1991): "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," NBER Working Paper No. 0115, November.
Angrist, J. D., E. Battistin, and D. Vuri (2017a): "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics,* **9**, 216–49.
Angrist, J. D., S. R. Cohodes, S. M. Dynarski, P. A. Pathak, and C. 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.
Angrist, J. D., S. M. Dynarski, T. J. Kane, P. A. Pathak, and C. R. Walters (2010a): "Inputs and Impacts in Charter Schools: KIPP Lynn," *American Economic Review: Papers & Proceedings,* **100**, 239–243.
— (2012): "Who Benefits from KIPP?" *Journal of Policy Analysis and Management,* **31**, 837–860.
Angrist, J. D. and W. N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review,* **88**, 450–477.
Angrist, J. D. and I. Ferna´ ndez-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," in *Advances in Economics and Econometrics,* ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge University Press, vol. **III** of *Econometric Society Monographs, Econometrics,* chap. 11, 401–434.
Angrist, J. D., K. Graddy, and G. 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.
Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017b): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, **132**, 871–919.
— (2021a): "Credible School Value-Added with Undersubscribed School Lotteries," MIT Blueprint Labs Working Paper. Forthcoming, *Review of Economics and Statistics.*
— (2021b): "Race and the Mismeasure of School Quality," NBER Working Paper No. 29608, December.
Angrist, J. D. and G. W. Imbens (1991): "Sources of Identifying Information in Evaluation Models," NBER Working Paper No. 0117, December.
— (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association,* **90**, 431–442.
Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association,* **91**, 444–455.
Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics,* **106**, 979–1014.
— (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.
Angrist, J. D. and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics,* **114**, 533–575.

Angrist, J. D., V. Lavy, J. Leder-Luis, and A. Shany (2019a): "Maimonides' Rule Redux," *American Economic Review: Insights,* **1**, 309–24.

Angrist, J. D., V. Lavy, and A. Schlosser (2010b): "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *Journal of Labor Economics,* **28**, 773–824.

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

Angrist, J. D., P. A. Pathak, and R. A. Za´ rate (2019b): "Choice and Consequence: Assessing Mismatch at Chicago Exam Schools," NBER Working Paper No. 26137, August.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion,* Princeton University Press.

— (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.

— (2014): *Mastering 'Metrics: The Path from Cause to Effect,* Princeton, Princeton University Press.

— (2017): "Undergraduate Econometrics Instruction: Through Our Classes, Darkly,"
Journal of Economic Perspectives, **31**, 125–44.

Angrist, J. D. and M. 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.

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

Ashenfelter, O. and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics,* **67**, 648–660.

Ashenfelter, O. and D. Card, eds. (1999): *The Handbook of Labor Economics,* vol. **3A**, Amsterdam, Elsevier.

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, **92**, 1171–1176.

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

Barrow, L., L. Sartain, and M. de la Torre (2020): "Increasing Access to Selective High Schools Through Place-Based Affirmative Action: Unintended Consequences," *American Economic Journal: Applied Economics,* **12**, 135–63.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," R*eview of Economic Studies,* **81**, 608–650.

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

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2017): "rdrobust: Software for Regression-Discontinuity Designs," *The Stata Journal,* **17**, 372–404.

Cattaneo, M. D., B. R. Frandsen, and R. 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.

Cattaneo, M. D., R. Titiunik, and G. 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.

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

Chamberlain, G. (1984): "Panel Data," in *The Handbook of Econometrics,* ed. by Z. Griliches and M. D. Intriligator, Amsterdam, Elsevier, vol. 2, 1247–1318.

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

Cook, T. 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.

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

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

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

Dehejia, R. H. and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, **94**, 1053–1062.

Dobbie, W. and R. G. Fryer, Jr (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.

Epstein, I. (1976): *Hebrew-English Translation of the Babylonian Talmud, Baba Bathra, Volume I,* London, Soncino Press.

Frandsen, B. R., L. J. Lefgren, and E. C. Leslie (2019): "Judging Judge Fixed Effects," NBER Working Paper No. 25528, February.

Frolich, M. and M. Huber (2019): "Including Covariates in the Regression Discontinuity Design," *Journal of Business and Economic Statistics,* **37**, 736–748.

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

Gerard, F., M. Rokkanen, and C. Rothe (2020): "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable," *Quantitative Economics,* **11**, 839–870.

Goldberger, A. S. (1972): "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," Working Paper.

Gronau, R. (1977): "Leisure, Home Production, and Work – the Theory of the Allocation of Time Revisited," *Journal of Political Economy,* **85**, 1099–1123.

Hahn, J. (1998): "On The Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, **66**, 315–331.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, **69**, 201–209.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature,* **24**, 1141–1177.

Hausman, J. A. (1983): "Specification and Estimation of Simultaneous Equation Models," in *The Handbook of Econometrics,* ed. by Z. Griliches and M. D. Intriligator, North- Holland, vol. **1**, 391–448.
Hearst, N., T. B. Newman, and S. B. Hulley (1986): "Delayed Effects of the Military Draft on Mortality," *New England Journal of Medicine,* **314**, 620–624.
Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, **73**, 669–738.
Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, **71**, 1161–1189.
Hoxby, C. M. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics,* **115**, 1239–1285.
Huber, M. and G. Mellace (2015): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, **97**, 398– 411.
Idoux, C. (2021): "Who Benefits from Selective School Attendance?" Working Paper.
Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, **62**, 467–475.
Imbens, G. W. and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies,* **79**, 933–959.
Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies,* **64**, 555–574.
Johnson, S. (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.
King, Jr, M. L. (1967): *Where Do We Go from Here: Chaos or Community?*, Boston, Beacon Press.
Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, **83**, 2043–2063.
Kolesár, M. and C. Rothe (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review,* **108**, 2277–2304.
Krueger, A. B. (1999): "Experimental Estimates of Education Production Functions,"
Quarterly Journal of Economics, **114**, 497–532.
LaLonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review,* 604–620.
Leamer, E. E. (1983): "Let's Take the Con Out of Econometrics," *American Economic Review,* **73**, 31–43.
Lee, D. S. and T. Lemieux (2010): "Regression Discontinuity Designs in Economics,"
Journal of Economic Literature, **48**, 281–355.
Mountjoy, J. and B. Hickman (2020): "The Returns to College(s): Estimating Value- Added and Match Effects in Higher Education," Becker Friedman Institute Working Paper.
Newey, W. K. (1985): "Generalized Method of Moments Specification Testing," *Journal of Econometrics,* 29, 229–256.
Newey, W. K. and K. D. West (1987): "Hypothesis Testing with Efficient Method of Moments Estimation," *International Economic Review,* **28**, 777–787.

Olson, L. and C. Jerald (2020): "The Big Test: The Future of Statewide Standardized Assessments," *FutureEd.*

Pinker, S. (2021): *Rationality: What It Is, Why It Seems Scarce, Why It Matters,* New York, Viking.

Robins, J. M. (2000): "Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models," in *Proceedings of the American Statistical Association,* Indianapolis, IN, vol. **1999**, 6–10.

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

Rosenzweig, M. R. and K. I. Wolpin (1980): "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, **48**, 227–240.

Sims, D. (2008): "A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California," *Journal of Policy Analysis and Management,* **27**, 457–478.

Snow, J. (1855): *On the Mode of Transmission of Cholera*, 2nd ed., London, Churchill.

Thernstrom, A. and S. Thernstrom (2004): No Excuses: *Closing the Racial Gap in Learning,* New York, Simon and Schuster.

Thistlethwaite, D. L. and D. T. Campbell (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology,* **51**, 309–317.

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