Chapter 3

Instrumental Variables

KWAI CHANG CAINE: From a single action, you draw an entire universe.

Kung Fu, Season 1, Episode 1

Our Path

Statistical control through regression may fail to produce convincing estimates of causal effects. Luckily, other paths lead to *other things equal*. Just as in randomized trials, the forces of nature, including human nature, sometimes manipulate treatment in a manner that obviates the need for controls. Such forces are rarely the only source of variation in treatment, but this is an obstacle easily surmounted. The *instrumental variables* (IV) method harnesses partial or incomplete random assignment, whether naturally occuring or generated by researchers. We illustrate this important idea three ways. The first evaluates an American education innovation—charter schools—with an elementary IV analysis that exploits randomized school admissions lotteries. A second IV application, examining the question of how best to respond to domestic violence, shows how IV can be used to analyze field experiments in which the subjects randomly assigned to treatment are free to opt out. The third application explores the long-run effects of growing up in a larger or smaller family. This application illustrates *two-stage least squares* (2SLS), an elaboration on the IV method and one of our most powerful tools.

3.1 The Charter Conundrum

INTERVIEWER: Have your mom and dad told you about the lottery?

DAISY: The lottery … isn't that when people play and they win money?

Waiting for Superman, 2010

The release of *Waiting for Superman,* a documentary film that tells the

story of applicants to charter schools in New York and California, intensified an already feverish debate over American education policy. *Superman* argues that charter schools offer the best hope for poor minority students who would otherwise remain at inner city public schools, where few excel and many drop out.

Charter schools are public schools that operate with considerably more autonomy than traditional American public schools. A charter—the right to operate a public school—is typically awarded to an independent operator (mostly private, nonprofit management organizations) for a limited period, subject to renewal conditional on good performance. Charter schools are free to structure their curricula and school environments. Many charter schools expand instruction time by running long school days and continuing school on weekends and during the summer. Perhaps the most important and surely the most controversial difference between charters and traditional public schools is 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. These contracts may improve working conditions for teachers, but they can make it hard to reward good teachers or dismiss bad ones.

Among the schools featured in *Waiting for Superman* is KIPP LA College Prep, one of more than 140 schools affiliated with the Knowledge Is Power Program. KIPP schools are emblematic of the No Excuses approach to public education, a widely replicated charter model that emphasizes discipline and comportment and features a long school day, an extended school year, selective teacher hiring, and a focus on traditional reading and math skills. KIPP was started in Houston and New York City in 1995 by veterans of Teach for America, a program that recruits thousands of recent graduates of America's most selective colleges and universities to teach in low-performing school districts. Today, the KIPP network serves a student body that is 95% black and Hispanic, with more than 80% of KIPP students poor enough to qualify for the federal government's subsidized lunch program.[1]

The American debate over education reform often focuses on the achievement gap, shorthand for uncomfortably large test score differences by race and ethnicity. Black and Hispanic children generally score well below white and Asian children on standardized tests. The question of how policymakers should react to large and persistent racial achievement gaps generates two sorts of responses. The first looks to schools to produce better outcomes; the second calls for broader social change, arguing that schools alone are unlikely to close achievement gaps. Because of its focus on minority students, KIPP is often central in this debate, with supporters pointing out that nonwhite KIPP students have markedly higher average test scores than nonwhite students from nearby schools. KIPP skeptics have argued that KIPP's apparent success reflects the fact that KIPP attracts families whose children are more likely to succeed anyway:

> KIPP students, as a group, enter KIPP with substantially higher achievement than the typical achievement of schools from which they came…. [T]eachers told us either that they referred students who were more able than their peers, or that the most motivated and educationally sophisticated parents were those likely to take the initiative … and enroll in KIPP.[2]

This claim raises the important question of whether *ceteris* is *paribus* when KIPP students are compared to other public school children.

Playing the Lottery

The first KIPP school in New England was a middle school in the town of Lynn, Massachusetts, just north of Boston. An old ditty warns: "Lynn, Lynn, city of sin, you never come out the way you came in." Alas, there's not much coming out of Lynn today, sinful or otherwise. Once a shoe manufacturing hub, Lynn has more recently been distinguished by high rates of unemployment, crime, and poverty. In 2009, more than three-quarters of Lynn's mostly nonwhite public school students were poor enough to qualify for a subsidized lunch. Poverty rates are even higher among KIPP Lynn's entering cohorts of fifth graders. Although urban charter schools typically enroll many poor, black students, KIPP Lynn is unusual among charters in enrolling a high proportion of Hispanic children with limited English

proficiency.

KIPP Lynn got off to a slow start when it opened in fall 2004, with fewer applicants than seats. A year later the school was oversubscribed, but not by much. After 2005, however, demand accelerated, with more than 200 students applying for about 90 seats in fifth grade each year. As required by Massachusetts law, scarce charter seats are allocated by lottery. More than a colorful institutional detail, these lotteries allow us to untangle the charter school causality conundrum. Our IV tool uses these admissions lotteries to frame a naturally occurring randomized trial.

The decision to attend a charter school is never entirely random: even among applicants, some of those offered a seat nevertheless choose to go elsewhere, while a few lottery losers find their way in by other means. However, comparisons of applicants who are and are not *offered* a seat as a result of random admissions lotteries should be satisfyingly apples to apples in nature. Assuming the only difference created by winning the lottery is in the likelihood of charter enrollment (an assumption called an *exclusion restriction*), IV turns randomized offer effects into causal estimates of the effect of charter attendance. Specifically, IV estimates capture causal effects on the sort of child who enrolls in KIPP when offered a seat in a lottery but wouldn't manage to get in otherwise. As we explain below, this group is known as the set of KIPP lottery *compliers*.

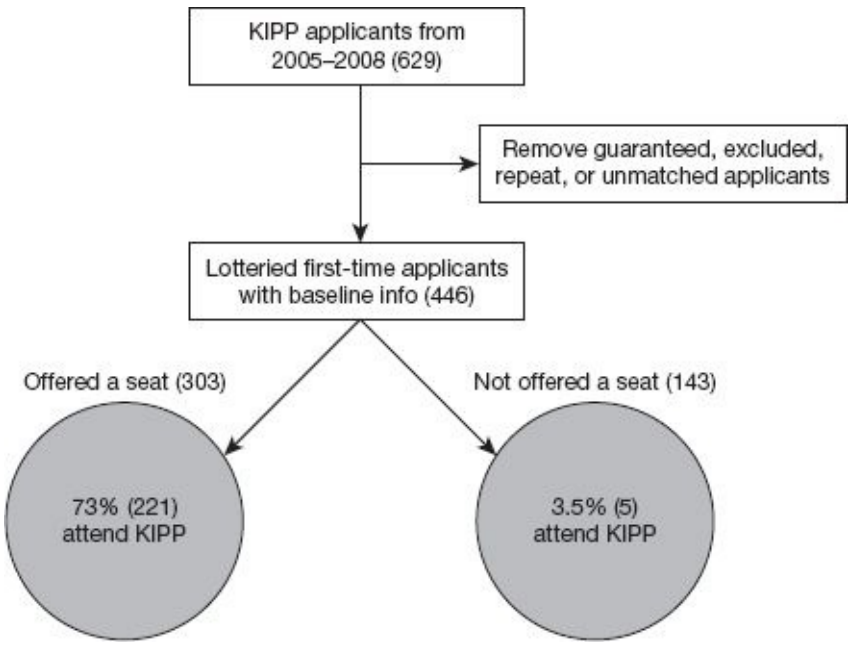
Master Joshway and his collaborators collected data on applicants to KIPP Lynn from fall 2005 through fall 2008.[3] Some applicants bypass the lottery: those with previously enrolled siblings are (for the most part) guaranteed admission. A few applicants are categorically excluded (those too old for middle school, for example). Among the 446 applicants for fifth-grade entry who were subject to random assignment in the four KIPP lotteries held from 2005 to 2008, 303 (68%) were offered a seat. Perhaps surprisingly, however, a fair number of these students failed to enroll come September. Some had moved away, while others ultimately preferred a nearby neighborhood school. Among those offered a seat, 221 (73%) appeared at KIPP the following school year. At the same time, a

handful of those not offered a place (about 3.5%) nevertheless found their way into KIPP (a few losing applicants were offered charter seats at a later date or in a later lottery). Figure 3.1 summarizes this important information.

KIPP lotteries randomize the offer of a charter seat. Random assignment of offers should balance the demographic characteristics of applicants who were and were not offered seats. Balance by offer status indeed looks good, as can be seen in panel A of Table 3.1. As a benchmark, the first column reports demographic characteristics and elementary school test scores for all Lynn public school fifth graders. The second and third columns, which report averages for KIPP lottery winners and the difference in means between winners and losers, show that winners and losers are about equally likely to be black or Hispanic or poor enough to qualify for a free lunch.

FIGURE 3.1

Application and enrollment data from KIPP Lynn lotteries



Note: Numbers of Knowledge Is Power Program (KIPP) applicants are shown in parentheses.

An especially important feature of Table 3.1 is the check for balance in pretreatment outcomes, namely, the test scores of lottery applicants in fourth grade, prior to KIPP enrollment (these are labeled "baseline scores" in the table). As is common in research on student

achievement, these scores have been *standardized* by subtracting the mean and dividing by the standard deviation of scores in a reference population, in this case, the population of Massachusetts fourth graders. After standardization, scores are measured in units defined by the standard deviation of the reference population. As in many poorer cities and towns in Massachusetts, average math scores in Lynn fall about three-tenths of a standard deviation below the state mean. This level of scores is written $-.3\sigma$ (as in the appendix to Chapters 1 and 2, standard deviation is represented by the Greek letter "sigma"). The small and statistically insignificant baseline differences between KIPP lottery winners and losers reported in column (3) of Table 3.1 are most likely due to chance.

TABLE 3.1
Analysis of KIPP lotteries

		KIPP applicants			
	Lynn public fifth graders (1)	KIPP Lynn lottery winners (2)	Winners vs. losers (3)	Attended KIPP (4)	Attended KIPP vs. others (5)
Panel A. Baseline characteristics					
Hispanic	.418	.510	−.058 (.058)	.539	.012 (.054)
Black	.173	.257	.026 (.047)	.240	−.001 (.043)
Female	.480	.494	−.008 (.059)	.495	−.009 (.055)
Free/Reduced price lunch	.770	.814	−.032 (.046)	.828	.011 (.042)
Baseline math score	−.307	−.290	.102 (.120)	−.289	.069 (.109)
Baseline verbal score	−.356	−.386	.063 (.125)	−.368	.088 (.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741 (.037)	1.000	1.000 —
Math score	−.363	−.003	.355 (.115)	.095	.467 (.103)
Verbal score	−.417	−.262	.113 (.122)	−.211	.211 (.109)
Sample size	3,964	253	371	204	371

Notes: This table describes baseline characteristics of Lynn fifth graders and reports

estimated offer effects for Knowledge Is Power Program (KIPP) Lynn applicants. Means appear in columns (1), (2), and (4). Column (3) shows differences between lottery winners and losers. These are coefficients from regressions that control for risk sets, namely, dummies for year and grade of application and the presence of a sibling applicant. Column (5) shows differences between KIPP students and applicants who did not attend KIPP. Standard errors are reported in parentheses.

The final two columns in Table 3.1 show averages for fifth graders who enrolled at KIPP Lynn, along with differences between KIPP applicants who did and did not enroll at KIPP. Since enrollment is not randomly assigned, differences between enrolled and nonenrolled students potentially reflect selection bias: Lottery winners who chose to go elsewhere may care less about school than those who accepted a KIPP enrollment opportunity. This is the selection bias scenario described by KIPP skeptics. As it turns out, however, the gaps in column (5) are small, and none approach statistical significance, suggesting that selection bias may not be important in this context after all.

Most KIPP applicants apply to enter KIPP in fifth grade, one year before regular middle school starts, but some apply to enter in sixth. We look here at effects of KIPP attendance on test scores for tests taken at the end of the grade following the application grade. These scores are from the end of fifth grade for those who applied to KIPP when they were in fourth grade and the end of sixth grade for those who applied to KIPP while in fifth. The resulting sample, which includes 371 applicants, omits young applicants who applied for entry after finishing third grade and a few applicants with missing baseline or outcome scores.[4]

Panel B of Table 3.1 shows that KIPP applicants who were offered a seat had standardized math scores close to 0, that is, near the state mean. Because KIPP applicants start with fourth-grade scores that average roughly .3σ below the state mean, achievement at the level of the state mean should be seen as impressive. By contrast, the average outcome score among those not offered a seat is about $-.36\sigma$, a little below the fourth-grade starting point.

Since lottery offers are randomly assigned, the difference between 0 and $-.36$, reported in column (3), is an average causal effect: the offer of a seat at KIPP Lynn boosts math scores by .36σ, a large gain

(the effect of KIPP offers on reading scores, though also positive, is smaller and not statistically significant). As a technical note, the analysis here is slightly more complicated than a simple comparison of means, though the idea is the same. The results in column (3) come from regressions of scores on a dummy variable indicating KIPP offers, along with dummies for year and grade of application and the presence of a sibling applicant. These control variables are necessary because the probability of winning the lottery varies from year to year and from grade to grade, and is much higher for siblings. The control variables used here describe groups of students (sometimes called *risk sets*) for whom the odds of a lottery offer are constant.[5]

What does an offer effect of .36σ tell us about the effects of KIPP Lynn attendance? The IV estimator converts KIPP offer effects into KIPP attendance effects. In this case, the *instrumental variable* (or "instrument" for short) is a dummy variable indicating KIPP applicants who receive offers. In general, an instrument meets three requirements:

(i) The instrument has a causal effect on the variable whose effects we're trying to capture, in this case KIPP enrollment. For reasons that will soon become clear, this causal effect is called the *first stage*.

(ii) The instrument is 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 family background or motivation). This is known as the *independence assumption*.

(iii) Finally, IV logic requires an *exclusion restriction*. The exclusion restriction describes a single channel through which the instrument affects outcomes. Here, the exclusion restriction amounts to the claim that the .36σ score differential between winners and losers is attributable solely to the .74 win-loss difference in attendance rates shown in column (3) of Table 3.1 (at the top of panel B).

The IV method uses these three assumptions to 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:

$$\begin{aligned}&\textit{Effect of offers on scores}\\&\quad=\big(\{\textit{Effect of offers on attendance}\}\\&\qquad\times\{\textit{Effect of attendance on scores}\}\big).\end{aligned}$$

Rearranging, the causal effect of KIPP attendance is

$$\begin{aligned}&\textit{Effect of attendance on scores}\\&\quad=\frac{\{\textit{Effect of offers on scores}\}}{\{\textit{Effect of offers on attendance}\}}.\end{aligned}\qquad(3.1)$$

This works out to be .48σ, as shown at the left in Figure 3.2.

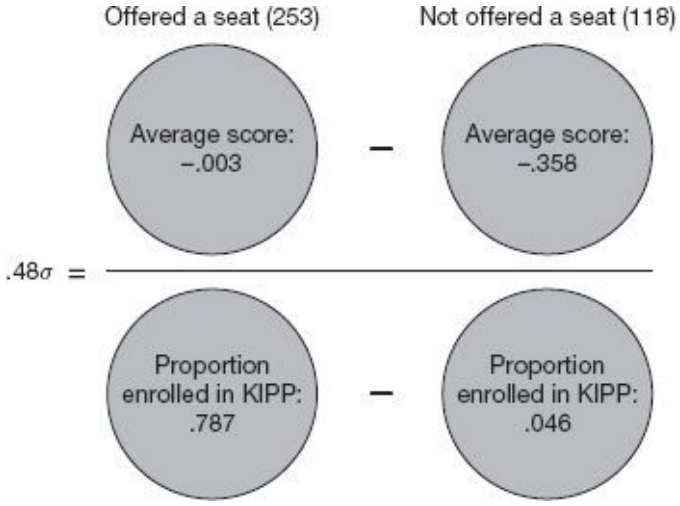
The logic generating equation (3.1) is easily summarized: KIPP offers are assumed to affect test scores via KIPP attendance alone. Offers increase attendance rates by about 75 percentage points (.74 to be precise), so multiplying effects of offers on scores by about 4/3 ($\approx$ 1/.74) generates the attendance effect. This adjustment corrects for the facts that roughly a quarter of those who were offered a seat at KIPP chose to go elsewhere, while a few of those not offered nevertheless wound up at KIPP.

An alternative estimate of the KIPP attendance effect appears in columns (4) and (5) in Table 3.1. Column (4) reports means for KIPP students, while column (5) shows the contrast between KIPP students and everyone else in the applicant pool. The differences in column (5) ignore randomized lottery offers and come from a regression of post-enrollment math scores on a dummy variable for KIPP attendance, along with the same controls used to construct the win/loss differences in column (3). The variation in KIPP attendance in this regression comes mostly, but not entirely, from the lottery. Because KIPP enrollment involves random assignment as well as individual choices (made, for example, when winners opt out), comparisons between those who do and don't enroll may be compromised by

selection bias. However, the estimate for math in column (5) (about .47σ) is close to the IV estimate in Figure 3.2, confirming our earlier conjecture that selection bias is unimportant in this case.

FIGURE 3.2

IV in school: the effect of KIPP attendance on math scores



Note: The effect of Knowledge Is Power Program (KIPP) enrollment described by this figure is .48σ = .355σ/.741.

A gain of half a standard deviation in math scores after one school year is a remarkable effect. Lynn residents lucky enough to have attended KIPP really don't come out the way they came in.

LATE for Charter School

The KIPP lottery exemplifies an IV chain reaction. The components of such reactions have been named, so masters can discuss them efficiently. We've noted that the original randomizer (in this case, a KIPP offer) is called an instrumental variable or just an instrument for short. As we've seen, the link from the instrument to the causal variable of interest (in this case, the effect of lottery offers on KIPP attendance) is called the first-stage, because this is the first link in the chain. The direct effect of the instrument on outcomes, which runs the full length of the chain (in this case, the effect of offers on scores), is called the *reduced form.* Finally, the causal effect of interest—the second link in the chain—is determined by the ratio of reduced form to first-stage estimates. This causal effect is called a *local average treatment effect* (*LATE* for short).

The links in the IV chain are made of differences between conditional expectations, that is, comparisons of population averages for different groups. In practice, population averages are estimated using sample means, usually with data from random samples. The necessary data are

- the *instrument,* Z_i: in this case, a dummy variable that equals 1 for applicants randomly offered a seat at KIPP (defined only for those participating in the lottery);
- the *treatment variable,* D_i: in this case, a dummy variable that equals 1 for those who attended KIPP (for historical reasons, this is sometimes called the endogenous variable); and
- the *outcome variable,* Y_i: in this case, fifth-grade math scores.

Key relationships between these variables, that is, the links in the IV chain, are parameters. We therefore christen them, you guessed it, in Greek.

THE FIRST STAGE $E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$; call this ϕ.

In the KIPP study, ϕ *("phi") is the difference in KIPP attendance rates between those who were and were not offered a seat in the lottery (equal to .74 in Figure 3.2).*

THE REDUCED FORM $E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$; call this ρ.

In the KIPP study, ρ *("rho") is the difference in average test scores between applicants who were and were not offered a seat in the lottery (equal to .36 in Figure 3.2).*

THE LOCAL AVERAGE TREATMENT EFFECT (LATE)

$$\lambda = \frac{\rho}{\phi} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}; \quad (3.2)$$

LATE, denoted here by λ ("lambda"), is the ratio of the reduced form to the first stage.

In the KIPP study, LATE is the difference in scores between winners and losers divided by the difference in KIPP attendance rates between winners and losers (equal to .48 in Figure 3.2).

We can estimate λ by replacing the four population expectations on the right-hand side of equation (3.2) with the corresponding sample averages, an estimator masters call IV. In practice, however, we usually opt for a method known as two-stage least squares (2SLS), detailed in Section 3.3 below. 2SLS implements the same idea, with added flexibility. Either way, the fact that parameters are estimated using samples requires us to quantify their sampling variance with the appropriate standard errors. It won't surprise you to learn that there's a formula for IV standard errors and that your econometric software knows it. Problem solved!

A more interesting question concerns the interpretation of λ: just who is LATE for charter school, you might ask. Children probably differ in the extent to which they benefit from KIPP. For some, perhaps those 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 0. For others, KIPP attendance may matter greatly. LATE is an average of these different individual causal effects. Specifically, LATE is the average causal effect for children whose KIPP enrollment status is determined solely by the KIPP lottery.

The biblical story of Passover explains that there are four types of children, and so it is with children today. We'll start with the first three types: Applicants like Alvaro are dying to go to KIPP; if they lose the lottery, their mothers get them into KIPP anyway. Applicants like Camila are happy to go to KIPP if they win, but stoically accept the verdict if they lose. Finally, applicants like Normando worry about long days and lots of homework. Normando doesn't really want to go to KIPP and refuses to do so when hearing that he has won a seat. Normando is called a *never-taker,* because his choice of school is unaffected by the lottery (it's the social worker who put his name in the hat). At the other end of KIPP kommitment, Alvaro is called an *always-taker.* He'll happily take a seat when offered, while his mother finds a way to make it happen for him even when he loses, perhaps by falsely claiming a sibling among the winners. For Alvaro, too, choice of school is unaffected by the lottery.

Camila attends KIPP when she wins the lottery but will regretfully take a seat in her neighborhood school if she loses (Camila's foster mother has her hands full; she wants the best for her daughter, but plays the hand she's dealt). Camila is the type of applicant who gives IV its power, because the instrument changes her treatment status. When her $Z_i = 0$, Camila's $D_i = 0$; and when her $Z_i = 1$, Camila's $D_i = 1$. IV strategies depend on applicants like Camila, who are called *compliers,* a group we indicate with the dummy variable, C_i. The term "compliers" comes from the world of randomized trials. In many randomized trials, such as those used to evaluate new drugs, the decision to comply with a randomized treatment assignment remains voluntary and nonrandom (experimental subjects who are randomly offered treatment may decline it, for example). Compliers in such trials are those who take treatment when randomly offered treatment but not otherwise. With lottery instruments, LATE is the average causal effect of KIPP attendance on Camila and other compliers who enroll at KIPP if and only if they win the lottery. IV methods are uninformative for always-takers like Alvaro and never-takers like Normando, because the instrument is unrelated to their treatment status.

TABLE 3.2

The four types of children

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

Note: KIPP = Knowledge Is Power Program.

Table 3.2 classifies children like Alvaro, Normando, and Camila, as well as a fourth type, called *defiers.* The columns indicate attendance choices made when $Z_i = 0$; rows indicate choices made when $Z_i = 1$. The table covers all possible scenarios for every applicant, not only

those we observe (for example, for applicants who won an offer, the table describes what they would have done had they lost). Never-takers like Normando and always-takers like Alvaro appear on the main diagonal. Win or lose, their choice of school is unchanged. At the bottom left, Camila complies with her lottery offer, attending KIPP if and only if she wins. The first stage, $E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$, is driven by such applicants, and LATE reflects average treatment effects in this group.

The defiers in Table 3.2 are those who enroll in KIPP only when *not* offered a seat in the lottery. The Bible refers to such rebels as "wicked," but we make no moral judgments. We note, however, that 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 might be 0 even if everyone benefits from KIPP attendance. Luckily, defiant behavior is unlikely in charter lotteries and many other IV settings. We therefore assume defiant behavior is rare to nonexistent. This no-defiers assumption is called *monotonicity,* meaning that the instrument pushes affected applicants in one direction only.

We've argued that instrumental variables can be understood as initiating a causal chain in which an instrument, Z_i, changes the variable of interest, D_i, in turn affecting outcomes, Y_i. The notion of a complier population tied to each instrument plays a key role in our interpretation of this chain reaction. The LATE theorem says that for any randomly assigned instrument with a nonzero first stage, satisfying both monotonicity and an exclusion restriction, the ratio of reduced form to first stage is LATE, the average causal effect of treatment on compliers.[6] Recall (from Section 1.1) that Y_{1i} denotes the outcome for i with the treatment switched on, while Y_{0i} is the outcome for the same person with treatment switched off. Using this notation and the parameters defined above, LATE can be written:

$$\lambda = \frac{\rho}{\phi} = E[Y_{1i} - Y_{0i}|C_i = 1].$$

Without stronger assumptions, such as a constant causal effect for

everybody (this is the model described by equation (1.3) in Chapter 1), LATE needn't describe causal effects on never-takers and always-takers.

It shouldn't surprise you that an instrumental variable is not necessarily helpful for learning about effects on people whose treatment status cannot be changed by manipulating the instrument. The good news here is that the population of compliers is a group we'd like to learn about. In the KIPP example, compliers are children likely to attend KIPP were the network to expand and offer additional seats in a lottery, perhaps as a consequence of opening a new school in the same area. In Massachusetts, where the number of charter seats is capped by law, the consequences of charter expansion is the education policy question of the day.

Researchers and policymakers are sometimes interested in average causal effects for the entire treated population, as well as in LATE. This average causal effect is called the *treatment effect on the treated* (TOT for short). TOT is written $E[Y_{1i} - Y_{0i}|D_i = 1]$. As a rule, there are two ways to be treated, that is, to have D_i switched on. One is to be treated regardless of whether the instrument is switched off or on. As we've discussed, this is the story of Alvaro, an always-taker. The remainder of the treated population consists of compliers who were randomly assigned $Z_i = 1$. In the KIPP study, the treated sample includes compliers who were offered a seat (like Camila) and always-takers (like Alvaro) who attend KIPP no matter what. The population of compliers who were randomly offered a seat is representative of the population of all compliers (including compliers who lose the lottery and go to public schools), but effects on always-takers need not be the same as effects on compliers. We might imagine, for example, that Alvaro is an always-taker because his mother senses that KIPP will change his life. The causal effect he experiences is therefore larger than that experienced by less-committed treated applicants, that is, by treated compliers.

Because the treated population includes always-takers, LATE and TOT are usually not the same. Moreover, neither of these average causal effects need be the same over time or in different settings (such

as at charter schools with fewer minority applicants). The question of whether a particular causal estimate has predictive value for times, places, and people beyond those represented in the study that produced it is called *external validity*. When assessing external validity, masters must ask themselves why a particular LATE estimate is big or small. It seems likely, for example, that KIPP boosts achievement because the KIPP recipe provides a structured educational environment in which many children—but perhaps not all—find it easy to learn. Children who are especially bright and independent might not thrive at KIPP. To explore the external validity of a particular LATE, we can use a single instrument to look at estimates for different types of students—say, those with higher or lower baseline scores. We can also look for additional instruments that affect different sorts of compliers, a theme taken up in Section 3.3. As with estimates from randomized trials, the best evidence for the external validity of IV estimates comes from comparisons of LATEs for the same or similar treatments across different populations.

3.2 Abuse Busters

The police were called to O. J. Simpson's Los Angeles mansion at least nine times over the course of his marriage to Nicole Brown Simpson. But the former National Football League superstar, nicknamed "The Juice," was arrested only once, in 1989, when he pleaded no contest to a charge of spousal abuse in an episode that put Nicole in the hospital. Simpson paid a small fine, did token community service, and was ordered to seek counseling from the psychiatrist of his choice. The prosecutor in the 1989 case, Robert Pingle, noted that Nicole had not been very cooperative with authorities in the aftermath of her severe beating. Five years later, Nicole Brown Simpson and her companion Ronald Goldman were murdered by an unknown intruder whom many believe was Nicole's ex-husband, O.J.[7]

How should police respond to domestic violence? Like Nicole Brown Simpson, abuse victims are often reluctant to press charges. Arresting batterers without victim cooperation may be pointless and could serve to aggravate an already bad situation. To many observers and not a few police officers, social service agencies seem best

equipped to respond to domestic violence. At the same time, victim advocates worry that the failure to arrest batterers signals social tolerance for violent acts that, if observed between strangers, would likely provoke a vigorous law enforcement response.

In the wake of a heated policy debate, the mayor and police chief of Minneapolis embarked on a pathbreaking experiment in the early 1980s. The Minneapolis Domestic Violence Experiment (MDVE) was designed to assess the value of arresting batterers.[8] The MDVE research design incorporated three treatments: arrest, ordering the suspected offender off the premises for 8 hours (separation), and a counseling intervention that might include mediation by the officers called to the scene (advice). The design called for one of these three treatments to be randomly selected whenever participating Minneapolis police officers encountered a situation meeting experimental criteria (specifically, probable cause to believe that a cohabitant or spouse had committed misdemeanor assault against a partner in the past 4 hours). Cases of life-threatening or severe injury (that is, felony assault) were excluded. Both suspect and victim had to be present at the time officers arrived. The primary outcome examined by the MDVE was the reoccurrence of a domestic assault at the same address within 6 months of the original random assignment.

The MDVE randomization device was a pad of report forms randomly color-coded for three possible responses: arrest, separation, and advice. Officers who encountered a situation that met experimental criteria were to act according to the color of the form on top of the pad. The police officers who participated in the experiment had volunteered to take part and were therefore expected to implement the research design. At the same time, everyone involved with the study understood that strict adherence to the randomization protocol was unrealistic and inappropriate.

TABLE 3.3

Assigned and delivered treatments in the MDVE

Assigned treatment	Delivered treatment			
		Coddled		
	Arrest	Advise	Separate	Total
Arrest	98.9 (91)	0.0 (0)	1.1 (1)	29.3 (92)
Advise	17.6 (19)	77.8 (84)	4.6 (5)	34.4 (108)
Separate	22.8 (26)	4.4 (5)	72.8 (83)	36.3 (114)
Total	43.4 (136)	28.3 (89)	28.3 (89)	100.0 (314)

Notes: This table shows percentages and counts for the distribution of assigned and delivered treatments in the Minneapolis Domestic Violence Experiment (MDVE). The first three columns show row percentages. The last column reports column percentages. The number of cases appears in parentheses.

In practice, officers often deviated from the responses called for by the color of the report form drawn at the time of an incident. In some cases, suspects were arrested even though random assignment called for separation or advice. Most arrests in these cases occurred when a suspect attempted to assault an officer, a victim persistently demanded an arrest, or when both parties were injured. A few deviations arose when officers forgot their report forms. As a result of these deviations from the experimental protocol, *treatment delivered* was not random. This can be seen in Table 3.3, which tabulates treatments assigned and delivered. Almost every case assigned to arrest resulted in arrest (91 of 92 cases assigned), but many cases assigned to the separation or advice treatments also resulted in arrest.

The contrast between arrest, which usually resulted in a night in jail, and gentler alternatives generates the most interesting and controversial findings in the MDVE. Table 3.3 therefore combines the two nonarrest treatments under the heading “coddled.” Random assignment had a large but not deterministic effect on the likelihood a suspected batterer was coddled: A case assigned to be coddled was coddled with probability .797 ($\frac{(84+5)+(5+83)}{108+114} = \frac{177}{222}$); while a case not assigned to coddling (that is, assigned to arrest) was coddled with probability .011 (1/92). Because coddling was not delivered randomly, the MDVE looks like a broken experiment. IV methods, however, readily fix it.

When LATE Is the Effect on the Treated

The LATE framework is motivated by an analogy between IV and

randomized trials. But some instrumental variables really come from randomized trials. IV methods allow us to capture the causal effect of treatment on the treated in spite of the nonrandom compliance decisions made by participants in experiments like the MDVE. In fact, the use of IV is usually necessary in such experiments. A naive analysis of the MDVE data based on treatment delivered is misleading.

Analysis of the MDVE based on treatment delivered is misleading because the cases in which police officers were supposed to coddle suspected batterers and actually did so are a nonrandom subset of all cases assigned to coddling. Comparisons of those who were and were not coddled are therefore contaminated by selection bias. Batterers who were arrested when assigned to coddling were often especially aggressive or agitated. Use of randomly assigned intention to treat as an instrumental variable for treatment delivered eliminates this source of selection bias.

As always, an IV chain reaction begins with the first stage.[9] The MDVE first stage is the difference between the probability of being coddled when assigned to be coddled and the probability of being coddled when assigned to be arrested. Let Z_i indicate assignment to coddling, and let D_i indicate incidents where coddling was delivered. The first stage for this setup is

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = .797 - .011 = .786,$$

a large gap, but still far from the difference of 1 we'd get if compliance had been perfect.

Unfortunately, domestic abuse is often a repeat offense, as can be seen in the fact the police were called for a second domestic violence intervention at 18% of the addresses in the MDVE sample. Most importantly from the point of view of MDVE researchers, recidivism was greater among suspects assigned to be coddled than among those assigned to be arrested. We learn this by calculating the effect of random assignment to coddling on an outcome variable, Y_i, that indicates at least one post-treatment episode of suspected abuse:

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = .211 - .097 = .114. \quad (3.3)$$

Given that the overall recidivism rate is 18%, this estimated difference of 11 percentage points is substantial.

In randomized trials with imperfect compliance, where treatment assigned differs from treatment delivered, effects of random assignment such as that calculated in equation (3.3) are called *intention-to-treat* (ITT) effects. An ITT analysis captures the causal effect of being assigned to treatment. But an ITT analysis ignores the fact that some of those assigned to be coddled were nevertheless arrested. Because the ITT effect does not take this noncompliance into account, it's too small relative to the average causal effect of coddling on those who were indeed coddled. This problem, however, is easily addressed: ITT effects divided by the difference in compliance rates between treatment and control groups capture the causal effect of coddling on compliers who were coddled as a result of the experiment.

Dividing ITT estimates from a randomized trial by the corresponding difference in compliance rates is another case of IV in action: We recognize ITT as the reduced form for a randomly assigned instrument, specifically, random assignment to coddling. As we've seen, many suspected batterers assigned to be coddled were nevertheless arrested. The regression of a dummy for having been coddled on a dummy for random assignment to coddling is the first stage that goes with this reduced form. The IV causal chain begins with random assignment to treatment, runs through treatment delivered, and ultimately affects outcomes.

The LATE estimate that emerges from the MDVE data is impressive: .114/.786 = .145, a large coddling effect, even in comparison with the corresponding ITT estimates. Remarkably, even though officers on the scene were highly selective in choosing whether to follow the experimental protocol, this estimate of LATE is likely to be a good measure of the causal effect of treatment delivered.

As always, the causal interpretation of LATE turns in part on the relevant exclusion restriction, which requires that the treatment variable of interest be the only channel through which the instrument affects outcomes. In the MDVE, the IV chain reaction begins with the

color of police officers' incident report forms. The exclusion restriction here requires that randomly assigned form color affect recidivism solely through the decision to arrest or to coddle suspected batterers. This seems like a reasonable assumption, all the more so as batterers and victims were unaware of their participation in an experimental study.

Are the modest complications of an IV analysis really necessary? Suppose we analyze the MDVE using information on treatment delivered, ignoring the nonrandom nature of decisions to comply with random assignment. The resulting analysis compares recidivism among those who were and were not coddled, with no further complications or adjustments:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = .216 - .129 = .087.$$

The estimated effect here is quite a bit smaller than the IV estimate of almost 15 percentage points.

Chapter 1 shows that without random assignment, comparisons of treated and untreated subjects equal the causal effect of interest plus selection bias. The selection bias that contaminates a naive analysis of the MDVE is the difference in potential recidivism (that is, in Y_{0i}) between batterers who were and were not coddled. Although much of the variation in coddling was produced by random assignment, officers on the scene also used discretion. Batterers who were arrested even though they'd been randomly assigned to be coddled were often especially violent or agitated, while suspects in cases where officers complied with a coddling assignment were typically more subdued. In other words, batterers who were coddled were less likely to abuse again in any case. The resulting selection bias leads the calculation based on treatment delivered to underestimate the impact of coddling. In contrast with the KIPP study (discussed in Section 3.1), selection bias matters here.

IV analysis of the MDVE eliminates selection bias, capturing average causal effects on compliers (in this case, the effect of coddling batterers in incidents in which officers were willing to comply with random assignment to coddling). An interesting and important feature

of the MDVE is the virtually one-sided nature of noncompliance in treatment delivered. When randomized to arrest, the police faithfully arrested (with only one exception in 92 cases). By contrast, more than 20% of those assigned to be coddled were nevertheless arrested.

The asymmetry in coddling compliance means there were almost no always-takers in the MDVE. In our IV analysis of the MDVE, always-takers are suspected batterers who were coddled without regard to treatment assigned. The size of this group is given by the probability of coddling when assigned to arrest, in this case, only 1/92. As we noted in Section 3.1, any treated population is the union of two groups, the set of compliers randomly assigned to be treated and the set of always-takers. With no always-takers, all of the treated are compliers, in which case, LATE is TOT:

$$\lambda = E[Y_{1i} - Y_{0i}|C_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1].$$

Applying the no-always-takers property to the MDVE, we see that LATE is the average causal effect of coddling on the coddled. Specifically, the TOT estimate emerging from the MDVE contrasts recidivism among the coddled ($E[Y_{1i}|D_i = 1]$) with the rates we would observe in a counterfactual world in which coddled batterers were arrested instead ($E[Y_{0i}|D_i = 1]$). This important simplification of the usual LATE story emerges in any IV analysis with no always-takers, including many other randomized trials with one-sided noncompliance. When some of those randomly assigned to treatment go untreated, but no one randomly assigned to the control group gets treated, IV methods using random intention to treat as an instrument for treatment delivered capture TOT.[10]

A final note on how much good 'metrics matters: It's hard to overstate the impact of the MDVE on U.S. law enforcement. Batterers in misdemeanor domestic assault cases are now routinely arrested. In many states, arrest in cases of suspected domestic abuse has become mandatory.

GRASSHOPPER: Master, the O.J. case came a decade after the MDVE. The pathbreaking MDVE research design did not save Nicole

Brown and Ron Goldman.

MASTER JOSHWAY: Social change happens slowly, Grasshopper. And the original MDVE analysts reported naive estimates based on treatment delivered, along with intention-to-treat effects. The IV estimates in my 2006 study are much larger.

GRASSHOPPER: Would Nicole and Ron have been saved if earlier analysts had used instrumental variables?

MASTER JOSHWAY: There are some things we can never know.

3.3 The Population Bomb

Population control or race to oblivion?

Paul Ehrlich, 1968

World population increased from 3 billion to 6 billion between 1960 and 1999, a doubling time of 39 years, and about half as long as the time it took to go from 1.5 billion to 3 billion. Only a dozen years passed before the seventh billion came along. But contemporary demographers agree that population growth has slowed dramatically. Projections using current fertility rates point to a doubling time of 100 years or more, perhaps even forever. One widely quoted estimate has population peaking at 9 billion in 2070.[11] Contemporary hand-wringing about sustainable growth notwithstanding, the population bomb has been defused—what a relief!

The question of how population growth affects living standards has both a macro side and a micro side. Macro demography traces its roots to the eighteenth-century English scholar Thomas Malthus, who argued that population size increases when food output increases, so much so that productivity gains fail to boost living standards. The unhappy Malthusian outcome is characterized by a permanent subsistence-level existence for most people. This pessimistic view of economic growth has repeatedly been falsified by history, but that hasn't prevented it from gaining traction among latter-day doomsayers. Biologist Paul Ehrlich's 1968 blockbuster *The Population Bomb* famously argued for a Malthusian scenario featuring imminent mass starvation in India. Since then, India's population has tripled, while Indian living standards have increased markedly.[12]

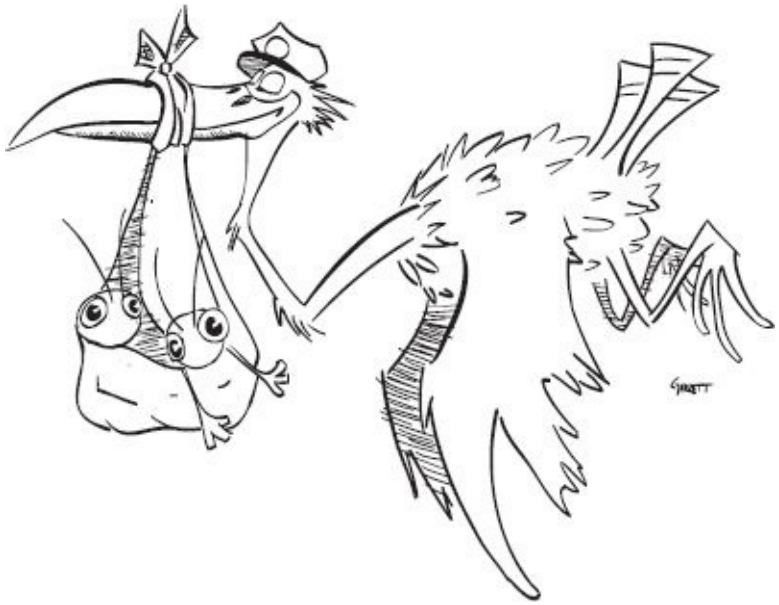
Economists have turned a micro lens on the relationship between family size and living standards. Here, attention focuses on the ability of households of different sizes to support a comfortable standard of living. We might indeed expect increases in family size to be associated with increased poverty and reduced education—more mouths to feed means less for each—and that's what simple correlations show. A more elaborate theoretical rationalization for this powerful relation comes from the work of the late Gary Becker and his collaborators. These studies introduced the notion of a "quantity-quality tradeoff," the idea that reductions in family size increase parental investment in children. For example, parents with fewer children might guard their children's health more closely and invest more in their schooling.[13]

On the policy side, the view that smaller families are essential for increasing living standards has motivated international agencies and many governments to promote, and occasionally even to require, smaller families. China led the way with the controversial One Child Policy, implemented in 1979. Other aggressive government-sponsored family planning efforts include a forced-sterilization program in India and the public promotion of family planning in Mexico and Indonesia. By 1990, 85% of people in the developing world lived in countries where the government considered high fertility to be a major force perpetuating poverty.[14]

The negative correlation between average family size and development indicators like schooling is hard to argue with. Is there a causal connection between family size and children's education? The challenge in answering this question, as always, is the *paribus*-ness of the *ceteris*. For the most part, fertility is determined by the choices parents make.[15] Not surprisingly, therefore, women with large families differ in many ways from those with smaller families; they tend to be less educated, for example. And the children of less-educated mothers tend to be less educated themselves. Marked differences in observable characteristics across families of different sizes raise the red flag of selection bias. Since women with different numbers of children are so observably different, we must acknowledge the possibility of important unobserved differences

associated with family size as well.

As always, the ideal solution to an omitted variables problem is random assignment. In this case, the experiment might go like this. (i) Draw a sample of families with one child. (ii) In some of these households, randomly distribute an additional child. (iii) Wait 20 years and collect data on the educational attainment of firstborns who did and did not get an extra sibling. Of course, we aren't likely to see such an experiment any time soon. Clever masters might, however, find sources of variation that reveal the causal connection between family size and schooling without the benefit of a real experiment.



Which brings us to the question of where babies come from. As most of our readers will know, human infants are delivered to households by a long-legged, long-necked bird called a stork (though it's a myth that the infant is dropped down the chimney—chimneys have a damper that prevents delivery of a live infant). Delivery occurs 9 months after a woman, whom we will refer to as the "mother," declares her intention to have a child. Storks are unresponsive to the wishes of men (except when these wishes are passed on by women), so we focus here on the notional experiment from the point of view of the mother and her oldest child.

The experiment we have in mind is the addition of children to households that have one already. The first-born child is our experimental subject. The 'metrics challenge is how to generate "as

good as randomly assigned" variation in family size for these subjects. Unfortunately, the Association of Stork Midwives rejects random assignment as unnatural. But storks nevertheless generate circumstantially random variation in family size by sometimes delivering more than one child in the form of twins (a consequence of the fact that storks are large and infants are small, so storks sometimes scoop multiples when picking babies in the infant storage warehouse). The fact that twins induce a family size experiment was first recognized in a pioneering study by Mark Rosenzweig and Kenneth Wolpin, who used a small sample of twins to investigate the quantity-quality trade-off in India.[16]

To exploit the twins experiment, we turn to a large sample from Israel, analyzed in a study of the quantity-quality tradeoff by Master Joshway, with colleagues Victor Lavy and Analia Schlosser (the "ALS study" for short).[17] Israel makes for an interesting case study because it has a very diverse population, including many people who were born in developing countries and into large families. About half of the Israeli Jewish population is of European ancestry, while the other half has roots in Asia or Africa. Quite a few Arabs live in Israel as well, but the data for Israeli non-Jews are less complete than for Jews. An attractive feature of the Israeli Jewish sample, besides ethnic diversity and larger families than are found in most developed countries, is the availability of information on respondents' families of origin, including the age and sex of their siblings. This unusual data structure is the foundation of the ALS empirical strategy.

We focus here on a group of first-born adults in a random sample of men and women born to mothers with at least two children. These firstborns have at least one younger sibling, but many have two or more. Consider a family in which the second birth is a singleton. On average, such families include 3.6 children. A second twin birth, however, increases average family size by .32, that is, by about one-third of a child. Why do twin births increase family size by a Solomonic fractional child? Many Israeli parents would like three or four children; their family size is largely unaffected by the occurrence of a multiple twin birth, since they were going to have more than two children either way. On the other hand, some families are happy with

only two children. The latter group is forced to increase family size from two to three when the stork delivers twins. The one-third-of-a-child twins differential in family size reflects a difference in probabilities: the likelihood of having a third child increases from about .7 with a singleton second birth to a certainty when the second birth is multiple. The .3 figure comes from the fact that the difference between a probability of 1 and probability of .7 is .3.

A simple regression of adult firstborns' highest grade completed on family size shows that each extra sibling is associated with a reduction of about one-quarter of a year of schooling (these results come from a model with age and sex controls). On the other hand, as the ALS study shows, even though first-born adults with second-born twin siblings were raised in larger families, they are no less educated than first-born adults in families where the second-born child was a singleton. The comparison of schooling between firstborns with twin and singleton siblings constitutes the reduced form for an IV estimate that uses twin births as an instrument for family size.

IV estimates are constructed from the ratio of reduced-form to first-stage estimates, so a reduced form of zero immediately suggests the causal effect of sibship size is also zero. The fact that the twins reduced-form and associated IV estimates are close to zero weighs against the view that a larger family of origin reduces children's schooling. In other words, the twins experiment generates no evidence of a quantity-quality tradeoff.

Multiple births have a marked effect on family size, but the twins experiment isn't perfect. Because the Association of Stork Midwives refuses to use random assignment, there's some imbalance in the incidence of twinning. Multiple births are more frequent among mothers who are older and for women in some racial and ethnic groups. This potentially leads to omitted variables bias in our analysis of the twins experiment, especially if some of the characteristics that boost twinning are hard to observe and control for.[18] Luckily, a second fertility experiment provides evidence on the quantity-quality trade-off.

In many countries, fertility is affected by sibling sex composition.

For one thing, parents often hope for a son; son preference is particularly strong in parts of Asia. In Europe, the Americas, and Israel, parents seem to care little about whether children are male or female. Rather, many parents hope for a diversified sibling-sex portfolio: Families whose first two children are both boys or both girls are more likely to have a third child. Because the sex of a newborn is essentially randomly assigned (male births occur about half the time and, in the absence of sex-selective abortion, little can be done to change this), parental preferences for mixed sibling-sex composition generate sex-mix instruments.

First-born Israeli adults who have a second-born sibling of the opposite sex grew up in households with about 3.60 children. But firstborns whose second-born sibling is of the same sex were raised in families with 3.68 children. In other words, the same-sex first stage for Israeli firstborns is about .08. As with the twins first stage, this differential reflects changes in the probability of childbearing induced by an instrument. In this case, the instrumental variable is a dummy variable that equals 1 for families whose first two children are both male or both female and equals 0 for families with one boy and one girl. While the sex-mix first stage is smaller than that arising from twinning, the number of families affected by same-sex sibships is much larger than the number of families affected by twinning. About half of all families with at least two children have either two boys or two girls at births number one and number two. By contrast, only about 1% of mothers have twins. Sibling sex composition also has a leg up on twinning in being unrelated to maternal characteristics, such as age at birth and race (as shown by ALS and in an earlier study by Master Joshway and William Evans).[19]

As it turns out, the educational attainment of first-born Israeli adults is unaffected by their siblings' sex composition. For example, the average highest grade completed by firstborns from families with mixed- and same-sex sibships is about equal at 12.6. Thus, the same-sex reduced form, and therefore the corresponding IV estimates, are both zero. Like the twins experiment, fertility changes generated by differences in sibling sex composition show no evidence of a quantity-quality trade-off.

The exclusion restriction required for a causal interpretation of sex-mix IV estimates asserts that sibling sex composition matters for adult outcomes only insofar as it changes family size. Might the sex-mix of the first two children affect children's educational outcomes for other reasons? Two boys and two girls are likely to share a bedroom longer than mixed-sex siblings, for example, and same-sex siblings may make better use of hand-me-down clothing. Such household efficiencies might make families with a same-sex sibship feel a little richer, a feeling that may ultimately increase parental investment in their children's schooling.

Can we test the exclusion restriction? Not directly, but, as is often the case, evidence can be brought to bear on the question. For some mothers, sex composition is unlikely to affect fertility. For example, in an Israeli sample, religious women who plan to have three or more children are always-takers for sex-mix instruments. On the other hand, highly educated women, most of whom plan small families, are never-takers if their fertility behavior is unchanged by sex mix. Because the fertility of always-takers and never-takers is unchanged by sibling sex composition, any relationship between sex-mix instruments and outcomes in samples with few compliers may signal violations of the underlying exclusion restriction.

We can express this idea more formally using the representation of LATE in equation (3.2). This expression defines LATE as the ratio of reduced-form to first-stage parameters, that is:

$$\lambda = \frac{\rho}{\phi},$$

which implies in turn that the reduced form, ρ, is the product of the first stage and LATE:

$$\rho = \phi\lambda.$$

From this we conclude that in samples where the first stage, ϕ, is zero, the reduced form should be zero as well. On the other hand, a statistically significant reduced-form estimate with no evidence of a corresponding first stage is cause for worry, because this suggests some channel other than the treatment variable (in this case, family

size) links instruments with outcomes. In this spirit, ALS identified demographic groups for which the effect of twins and sex-composition instruments on family size is small and not significantly different from zero. These "no-first-stage samples" generate no evidence of significant reduced-form effects that might signal violations of the exclusion restriction.

One-Stop Shopping with Two-Stage Least Squares

IV estimates of causal effects boil down to reduced-form comparisons across groups defined by the instrument, scaled by the appropriate first stage. This is a universal IV principle, but the details vary across applications. The quantity-quality scenario differs from the KIPP story in that we have more than one instrument for the same underlying causal relation. Assuming that twins and sex-mix instruments both satisfy the required assumptions and capture similar average causal effects, we'd like to combine the two IV estimates they generate to increase statistical precision. At the same time, twinning might be correlated with maternal characteristics like age at birth and ethnicity, leading to bias in twins IV estimates. We'd therefore like a simple IV procedure that controls for maternal age and any other confounding factors. This suggests a payoff to integrating the IV idea with the regression methods discussed in Chapter 2.

Two-stage least squares (2SLS) generalizes IV in two ways. First, 2SLS estimates use multiple instruments efficiently. Second, 2SLS estimates control for covariates, thereby mitigating OVB from imperfect instruments. To see how 2SLS works, it helps to rewrite the first stage (ϕ) and reduced form (ρ) parameters as regression coefficients instead of differences in means. Starting with a single instrument, say, a dummy variable for multiple second births denoted by Z_i, the reduced-form effect can be written as the coefficient ρ in the regression equation:

$$Y_i = \alpha_0 + \rho Z_i + e_{0i}. \qquad (3.4)$$

As we noted in the appendix to Chapter 2, regression on a constant term and a single dummy variable produces the difference in the conditional means of the dependent variable with the dummy

switched off and on. The coefficient on Z_i in equation (3.4) is therefore

$$\rho = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0].$$

Likewise, the first-stage effect of Z_i is the coefficient ϕ in the first-stage equation:

$$D_i = \alpha_1 + \phi Z_i + e_{1i}, \qquad (3.5)$$

where $\phi = E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$. Since $\lambda = \rho/\phi$, we conclude that LATE is the ratio of the slope coefficients in regressions (3.4) and (3.5).

The 2SLS procedure offers an alternative way of computing ρ/ϕ. The 2SLS name comes from the fact that LATE can be obtained from a sequence of two regressions. In the 2SLS first stage, we estimate equation (3.5) and save the fitted values, $\hat{D}_i$. These "first-stage fits" are defined as

$$\hat{D}_i = \alpha_1 + \phi Z_i. \qquad (3.6)$$

The 2SLS second stage regresses Y_i on $\hat{D}_i$, as in

$$Y_i = \alpha_2 + \lambda_{2SLS}\hat{D}_i + e_{2i}.$$

The value of λ_{2SLS} generated by this second step is identical to the ratio of reduced form to first-stage regression coefficients, ρ/ϕ, a theoretical relationship derived in the chapter appendix.

Control variables like maternal age fit neatly into this two-step regression framework.[20] Adding maternal age, denoted A_i, the reduced form and first stage look like

$$\text{Reduced form:} \quad Y_i = \alpha_0 + \rho Z_i + \gamma_0 A_i + e_{0i} \qquad (3.7)$$

$$\text{First stage:} \quad D_i = \alpha_1 + \phi Z_i + \gamma_1 A_i + e_{1i}. \qquad (3.8)$$

Here, the first-stage fitted values come from models that include the control variable, A_i:

$$\hat{D}_i = \alpha_1 + \phi Z_i + \gamma_1 A_i.$$

2SLS estimates are again constructed by regressing Y_i on both $\hat{D}_i$ and A_i. Hence, the 2SLS second-stage equation is

$$Y_i = \alpha_2 + \lambda_{2SLS}\hat{D}_i + \gamma_2 A_i + e_{2i}, \qquad (3.9)$$

which also includes A_i.

The 2SLS setup allows as many control variables as you like, provided they appear in both the first and second stages. As discussed in the chapter appendix, the corresponding covariate-adjusted LATE can still be constructed from the ratio of reduced-form to first-stage coefficients, ρ/ϕ. Indeed, we should separately inspect the upstairs and downstairs in this ratio to make sure all on both floors is kosher. But when it comes time to report results to the public, 2SLS is the way to go even in relatively simple scenarios like this one. Econometrics software packages compute 2SLS estimates directly, reducing the scope for mistakes and generating appropriate standard errors at no extra charge.[21]

What about our second family-size instrument, a dummy for same-sex sibships? Call this W_i (where $W_i = 1$ indicates two girls or two boys, and $W_i = 0$ otherwise). Here, too, control variables are called for, in particular, the sex of the first-born, which we code as a dummy, B_i, indicating first-born boys (as a rule, boys are born slightly more often than girls, so the probability of a same-sex pair is slightly higher when the firstborn is male). With two instruments, W_i and Z_i, and the extra control variable, B_i, the 2SLS first stage becomes

$$D_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma_1 A_i + \delta_1 B_i + e_{1i}. \quad (3.10)$$

The first-stage effects of the twins and sex-mix instruments are distinguished by subscripts t for twins and s for sex-mix: we write these as ϕ_t and ϕ_s. Both instruments appear with similarly subscripted coefficients in the corresponding reduced form as well:

$$Y_i = \alpha_0 + \rho_t Z_i + \rho_s W_i + \gamma_0 A_i + \delta_0 B_i + e_{0i}.$$

With these ingredients at hand, it's time to cook!

Second-stage estimates with two instruments and two covariates are

generated by the regression equation

$$Y_i = \alpha_2 + \lambda_{2SLS}\hat{D}_i + \gamma_2 A_i + \delta_2 B_i + e_{2i}, \quad (3.11)$$

where the fitted values, $\hat{D}_i$, come from first-stage equation (3.10). Note that the covariates appear at every turn: in the first and second stages, and in the reduced form. Equation (3.11) produces a weighted average of the estimates we'd get using the instruments Z_i and W_i one at a time, while controlling for covariates A_i and B_i. When the instruments generate similar results when used one at a time, the 2SLS weighted average is typically a more precise estimate of this common causal effect.

TABLE 3.4

Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320	.437			.449
	(.052)	(.050)			(.050)
Same-sex sibships			.079	.073	.076
			(.012)	(.010)	(.010)
Male		−.018		−.020	−.020
		(.010)		(.010)	(.010)
Controls	No	Yes	No	Yes	Yes

Notes: This table reports coefficients from a regression of the number of children on instruments and covariates. The sample size is 89,445. Standard errors are reported in parentheses.

2SLS offers a wonderfully flexible framework for IV estimation. In addition to incorporating control variables and using multiple instruments efficiently, the framework accommodates instruments of all shapes and sizes, not just dummy variables. In practice, however, masters use special-purpose statistical software to calculate 2SLS estimates instead of estimating regressions on fitted values like (3.11). Estimation of this equation, known as "manual 2SLS," doesn't produce the correct standard errors needed to measure sampling variance. The chapter appendix explains why.

Estimates of twins and sex-mix first stages with and without covariates appear in Table 3.4. The estimate from a first-stage model with controls, reported in column (2) of the table, shows that first-born Israeli adults whose second-born siblings were twin were raised in families with about .44 more children than those raised in families where the second birth was a singleton. This first-stage estimate is larger than the estimate of .32 computed without controls (reported in column (1)). The OVB formula therefore tells us that twin births are associated with factors that reduce family size, like older maternal age. Adjusting for maternal age and other possible confounding factors boosts the twins first stage. On the other hand, the same-sex first stage of .073 generated by a model with covariates is close to the uncontrolled estimate of .079, since sex mix is essentially unrelated to the included controls (these estimates can be seen in columns (3) and (4)). The fact that the first-born is male also has little effect on the size of his family. This can be seen in the small, marginally significant male coefficients reported in the last row (this is the only covariate coefficient reported in the table, though the presence of other controls is indicated in the bottom row).[22]

Second-stage estimates of the quantity-quality trade-off are reported in Table 3.5, along with the corresponding estimates from a conventional (that is, uninstrumented) OLS regression of the form

$$Y_i = \alpha_3 + \beta D_i + \gamma_3 A_i + \delta_3 B_i + e_{3i}.$$

The conventional regression estimates in column (1) show a strong negative relation between family size and education outcomes, even after adjusting for family background variables related to ethnicity and mother's age at birth. By contrast, the 2SLS estimates generated by twins instruments, reported in column (2) of the table, mostly go the other way, though the 2SLS estimates in this case are not significantly different from zero. Estimation using sex-composition instruments reinforces the twins findings. The 2SLS estimates in column (3) show uniformly positive effects of family size on education (though only one of these is significantly different from zero).

TABLE 3.5

OLS and 2SLS estimates of the quantity-quality trade-off

Dependent variable	OLS estimates (1)	2SLS estimates: Twins instruments (2)	2SLS estimates: Same-sex instruments (3)	2SLS estimates: Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in column (1). Columns (2), (3), and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Sample sizes are 89,445 for rows (1) and (2); 50,561 for row (3); and 50,535 for row (4). Standard errors are reported in parentheses.

An important feature of both the twins and sex-composition second stages is their precision, or lack thereof. IV methods discard all variation in fertility except that generated by the instrument. This can leave too little variation for statistically conclusive findings. We can increase precision, however, by pooling multiple instruments, especially if, when taken one at a time, the instruments generate similar findings (in this case, both twins and sex-composition instruments show little evidence of a quantity-quality trade-off). The resulting pooled first-stage estimates appear in column (5) of Table 3.4, while the corresponding second-stage results are reported in column (4) of Table 3.5.

The pooled second-stage estimates are not very different from those generated using the instruments one at a time, but the standard errors are appreciably smaller. For example, the estimated effect of family size on highest grade completed using both instruments is .24, with a standard error of .13, a marked drop from the standard errors of about .17 and .21 using twins and same-sex instruments one at a time. Importantly, the regression estimate in column (1), a very precise −.15 for highest grade completed, lies well outside the confidence interval associated with the 2SLS estimate in column (4).[23] This suggests that the strong negative association between family size and schooling is driven in large part and perhaps entirely by selection

bias.

MASTER JOSHWAY: Build the house of IV, Grasshopper.

GRASSHOPPER: The foundation has three layers: (i) the *first-stage* requires instruments that affect the causal channel of interest; (ii) the *independence assumption* requires instruments to be as good as randomly assigned; (iii) the *exclusion restriction* asserts that a single causal channel connects instruments with outcomes.

MASTER JOSHWAY: Can these assumptions be checked?

GRASSHOPPER: Check the first stage by looking for a strong relationship between instruments and the proposed causal channel; check independence by checking covariate balance with the instrument switched off and on, as in a randomized trial.

MASTER JOSHWAY: And exclusion?

GRASSHOPPER: The exclusion restriction is not easily verified. Sometimes, however, we may find a sample where the first stage is very small. Exclusion implies such samples should generate small reduced-form estimates, since the hypothesized causal channel is absent.

MASTER JOSHWAY: How are IV estimates computed?

GRASSHOPPER: Statistical software computes two-stage least squares estimates for us. This allows us to add covariates and use more than one instrument at a time. But we look at the first-stage and reduced-form estimates as well.

Masters of 'Metrics: The Remarkable Wrights

The IV method was invented by economist Philip G. Wright, assisted by his son, Sewall, a geneticist. Philip wrote frequently about agricultural markets. In 1928, he published *The Tariff on Animal and Vegetable Oils.*[24] Most of this book is concerned with the question of whether the steep tariffs on farm products imposed in the early 1920s benefited domestic producers. A 1929 reviewer noted that "Whatever the practical value of the intricate computation of elasticity of demand and supply as applied particularly to butter in this chapter, the discussion has high theoretical value."[25]