TECHNICAL WORKING PAPER SERIES

INSTRUMENTAL VARIABLES METHODS IN EXPERIMENTAL CRIMINOLOGICAL RESEARCH: WHAT, WHY, AND HOW?

Joshua Angrist

Technical Working Paper 314
http://www.nber.org/papers/T0314

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2005

Special thanks to Richard Berk, Howard Bloom, David Weisburd, and the participants in the May 2005 Jerry Lee Conference for the stimulating discussions that led to this paper. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Instrumental Variables Methods in Experimental Criminological Research: What, Why, and How?

Joshua Angrist

NBER Technical Working Paper No. 314
September 2005
JEL No. C12, C30

ABSTRACT

Quantitative criminology focuses on straightforward causal questions that are ideally addressed with randomized experiments. In practice, however, traditional randomized trials are difficult to implement in the untidy world of criminal justice. Even when randomized trials are implemented, not everyone is treated as intended and some control subjects may obtain experimental services. Treatments may also be more complicated than a simple yes/no coding can capture. This paper argues that the instrumental variables methods (IV) used by economists to solve omitted variables bias problems in observational studies also solve the major statistical problems that arise in imperfect criminological experiments. In general, IV methods estimate causal effects on subjects who comply with a randomly assigned treatment. The use of IV in criminology is illustrated through a re-analysis of the Minneapolis Domestic Violence Experiment.

Joshua Angrist
MIT Department of Economics
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
angrist@mit.edu

Background

I'm not a criminologist, but I've long admired criminology from afar. As an applied economist who puts the task of convincingly answering straightforward causal questions at the top of my agenda, I've been impressed with the no-nonsense outcomes-oriented approach taken by many quantitative criminologists. Does capital punishment deter? Do drug courts reduce recidivism? Does arrest for domestic assault reduce the likelihood of a repeat offense? These are the sort of important and straightforward causal questions that I can imagine studying myself.

I also appreciate the focus on credible research designs reflected in much of the criminological research agenda. Especially noteworthy is the fact that, in marked contrast with an unfortunate trend in education research, criminologists do not appear to have been afflicted with what psychologist Tom Cook (2001) calls "sciencephobia." This is a tendency to eschew rigorous quantitative research designs in favor of a softer approach that emphasizes process over outcomes. In fact, of the disciplines tracked in a survey of social science research methods by Boruch, de Moya, and Snyder (2002), Criminology is the only one to show a marked *increase* in the use of randomized trials since the mid-sixties.

The use of randomized trials in criminology is clearly increasing and, by now, experiments have been used to study interventions in policing, prevention, corrections, and courtrooms (Farrington and Welsh, 2005). Randomized trials are increasingly seen as the gold standard for scientific evidence in the crime field, as they are in medicine (Weisburd, *et al*, 2001). At the same time, a number of considerations appear to limit the applicability of randomized research designs to criminology.

A major concern in the criminological literature is the possibility of a failed research design (see, e.g., Farrington, 1983, Rezmovic, *et al*, 1981, and Gartin, 1995). Gartin (1995) notes that two sorts of design failure seem especially likely. The first, *treatment dilution*, is when subjects or units assigned to the treatment group do not get treated. The second, *treatment migration*, is when subjects or units in the control group nevertheless obtain the experimental treatment. These scenarios are indeed potential threats

to the validity of a randomized trial. For one thing, with non-random crossovers, the group actually treated may no longer be comparable to the group that ends up untreated. In addition, if intended treatment is only an imperfect proxy for treatment received, it seems clear that an analysis based on the original intention-to-treat must understate the causal effect of treatment *per se*.[2]

The purpose of this paper is to show how the instrumental variables (IV) methods widely used in Economics solve both the treatment dilution and treatment migration problems. As a by-product, the IV framework also opens up the possibility of a much wider range of flexible experimental research designs. These designs are less likely to raise the sort of ethical questions that are seen as limiting the applicability of traditional experimental designs in crime and justice (see, e.g., Weisburd, 2003, for a discussion). Finally, the logic of IV suggests a number of promising quasi-experimental research designs that might provide a reasonably credible (and inexpensive) substitute for an investigator's own random assignment.[3]

Motivation: The Minneapolis Domestic Violence Experiment

Treatment migration and treatment dilution are features of one of the most influential randomized trials in criminological research, the Minneapolis domestic violence experiment (MDVE), discussed in Sherman and Berk (1984) and Berk and Sherman (1988). The MDVE was motivating by debate over the importance of deterrence effects in the police response to domestic violence. Police are often reluctant to make arrests for domestic violence unless the victim demands an arrest, or the suspect does something that warrants arrest (beside the assault itself). As noted by Berk and Sherman (1988), this attitude has

---

[2]The problem of deviations from random assignment is not unique to criminology. Social experiments in labor economics often allow those selected for treatment to opt out (an example is the Illinois unemployment insurance bonuses experiment; see Woodbury and Spiegelman, 1987). Even in double-blind clinical trials, clinicians sometimes decipher and change treatment assignments (Schultz, 1995).

[3]The brief discussion in this paper glosses over a number of technical details. For a more comprehensive introduction to IV see Angrist and Krueger (2001, 1999), or the chapters on IV in Wooldridge (2003).

many sources: a general reluctance to intervene in a family disputes, the fact that domestic violence cases may not be prosecuted, genuine uncertainty as to what the best course of action is, and an incorrect perception that domestic assault cases are especially dangerous for arresting officers.

In response to a politically charged policy debate as to the wisdom of making arrests in response to domestic violence, the MDVE was conceived as a social experiment that might provide a resolution. The research design incorporated three treatments: arrest, ordering the offender off the premises for 8 hours, and some form of advice that might include mediation. The research design called for one of these three treatments to be randomly selected each time participating Minneapolis police officers encountered a situation meeting the experimental criteria (some kind of apparent misdemeanor domestic assault where there was probable cause to believe that a cohabitant or spouse had committed an assault against the other party in the past 4 hours). Cases of life-threatening or severe injury, i.e., felony assault, were excluded. Both suspect and victim had to be present upon the officer's arrival.

The randomization device was a pad of report forms that were randomly color-coded for each of the three possible response. Officers who encountered a situation that met the 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 comply with the research design. On the other hand, deviations from random assignment were allowed and even anticipated by the experimenters.

In practice, officers often deviated from the response called for by the color of the report form drawn at the time of an incident. In some cases, suspects were arrested when random assignment called for separation or advice. Officers would arrest in these cases when a suspect attempted to assault an officer, a victim persistently demanded an arrest, or if both parties were injured. In one case where random assignment called for arrest, officers separated instead. In a few cases, advice was swapped for separation and vice versa. Although most deviations from the intended treatment reflected purposeful

action on the part of the officers involved, sometimes deviations arose when officers simply forgot to bring their report forms.

Non-compliance with random assignment is not unique to the MDVE or criminological research. Any experimental intervention where ethical or practical considerations lead to a deviation from protocol is likely to have this feature. In practice, non-compliance is usually avoidable in research using human subjects. Gartin (1995) discusses criminological examples; non-compliance has long been discussed in randomized clinical trials (see, e.g., Efron and Feldman, 1991).

In the MDVE, the most common deviation from random assignment was the failure to separate or advise when random assignment called for this. This can be seen in Table 1, taken from Sherman and Berk (1984), which reports a cross-tabulation of treatment assigned and treatment delivered. Of the 92 suspects randomly assigned to be arrested, 91 were arrested. In contrast, of the 108 suspects randomly assigned to receive advice, 19 were arrested and 5 were separated. The compliance rate with the advice treatment was 78 percent. Likewise, of the 114 suspects randomly assigned to be separated 26 were arrested and 5 were advised. The compliance rate with the separation treatment was 73 percent.

The random assignment of *intended* treatments in the MDVE does not appear to have been subverted (Berk and Sherman, 1988). At the same time, it's clear that delivered treatments had a substantial behavioral component. Treatment delivered was, in the language of econometrics, *endogenous*. In other words, delivered treatments were determined in part by unobserved features of the situation that were very likely correlated with outcome variables such as re-offense. For example, some of the suspects who were arrested in spite of having been randomly assigned to receive advice or be separated were especially violent. An analysis that contrasts outcomes according to the treatment

delivered is therefore likely to be misleading, generating an over-estimate of the power of advice or separation to deter violence. I show below that this is indeed the case.[4]

A simple, commonly-used approach to the analysis of randomized clinical trials with imperfect compliance is to compare subjects according to original random assignment, ignoring compliance. This is known as an intention-to-treat (ITT) analysis. Because ITT comparisons use only the original random assignment, and ignore information on treatments actually delivered, they indeed provide unbiased estimates of the causal effect of researchers' intention-to-treat. At the same time, ITT estimates are almost always too small relative to the causal effect of interest, the effect of treatment itself.

An easy way to see why ITT is "too small" is to consider the ITT effect generated by an experiment where the likelihood of treatment turns out to be the same in both intended-treatment and intended-control groups. In this case, there is essentially "no experiment," i.e., the treatment-intended group is treated, on average, just like the control group. The resulting ITT effect is therefore zero, even though the causal effect of treatment on individuals may be positive or negative for everyone. More generally, the ITT effect is diluted by non-compliance. Thus, ITT provides a poor predictor of the average causal effect of similar interventions in the future, should future compliance rates differ.

A third strategy for dealing with compliance problems is to try to model the compliance decision, and somehow bring this model into the analysis of experimental data. Examples include the Berk, Smyth, and Sherman (1988) analysis of MDVE and the analysis in Efron and Feldman (1991). Except under strong and probably unrealistic assumptions, however, the noncompliance problem cannot be resolved by behavioral modeling (or by conditioning on the predicted compliance rates generated by a behavioral model). If it could, then we wouldn't need random assignment in the first place, since decisions to take

---

[4]The fact that those who comply with randomly assigned treatments are special can be seen in medical trials, where those who comply with protocol by taking a randomly assigned experimental treatment with no

treatment in a randomized experiment are no less endogenous than the decision to take treatment in an observational study. Luckily, however, behavioral models of the compliance process are unnecessary.

The Instrumental-Variables framework

The simplest and most robust solution to the treatment-dilution and treatment-migration problems is instrumental variables. This can be seen most easily using a conceptual framework that postulates a set of potential outcomes that could be observed in alternative states of the world. Originally introduced by statisticians in the 1920s as a way to discuss treatment effects in randomized experiments, the potential-outcomes framework has become the conceptual workhouse for non-experimental as well as experimental studies in medicine and social science (see, Holland, 1986, for a survey and Rubin, 1974 and 1977, for influential early contributions).

To link the abstract discussion to the MDVE example, I'll start with an interpretation of the MDVE as randomly assigning and delivering a single treatment. Because the policy discussion in the domestic assault context focuses primarily on the decision to arrest and possible alternatives, I define a binary (dummy) treatment variable for not arresting, which I'll call *coddling*. A suspect was randomly assigned to be coddled if the officer on the scene was instructed by the random assignment protocol (i.e., the color-coded report forms) to advise or separate. A subject received the coddling treatment if the treatment delivered was advice or separation. Later, I'll outline an IV setup for the MDVE that allows for multiple treatments.

The most important outcome variable in the MDVE was re-offense, i.e., the occurrence of post-treatment domestic assault by the same suspect. Let Y_i denote the observed re-offense status of suspect *i*. The potential outcomes in the binary-treatment version of MDVE are the re-offense status of suspect *i* if

---

clinical effects – i.e., a placebo – are often healthier than those who don't (as in the study analyzed by Efron and Feldman, 1991).

he were coddled, denotedY_{1i}, and the re-offense status of suspect *i* if he were not coddled, denotedY_{0i}. Both of these potential outcomes are assumed to be well-defined for each suspect even though only one is ever observed. Let D_i note treatment status based on treatment delivered. Then we can write the observed outcome variable as

$$Y_i = Y_{0i}(1-D_i) + Y_{1i}D_i.$$

A natural place to start any empirical analysis is by comparing outcomes on the basis of treatment delivered. Because of the non-random-nature of treatment delivery, however, such naive comparisons are likely to be misleading. This can be seen formally by writing

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

The first term in this decomposition is the average causal effect of treatment on the treated (ATET), a parameter of primary interest in evaluation research. ATET tells us the difference between average outcomes for the treated, $E[Y_{1i}| D_i=1]$, and what would have happened to treated subjects if they had not been treated, $E[Y_{0i}|D_i=1]$. The second term in (1) is the selection bias induced by the fact that treatment delivered was not randomly assigned. In the MDVE, those coddled were probably less likely to re-offend even in the absence of treatment. Hence $\{E[Y_{0i}|D_i=1] - E[Y_{0i}|D_i=0]\}$ is probably negative.

Selection bias disappears when delivered treatment is determined in a manner independent of potential outcomes, as in a randomized trial with perfect compliance. We then have

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

With perfect compliance, the simple treatment-control comparison recovers ATET. Moreover, because $\{Y_{1i},Y_{0i}\}$ is assumed to be independent of D_i in this case, ATET is also the population average treatment effect, $E[Y_{1i}-Y_{0i}]$.

The most important consequence of non-compliance is the likelihood of a relation between potential outcomes and delivered treatments. This relation confounds analyses based on delivered

treatments because of the resulting selection bias. But we have an ace in the hole: the compliance problem does not compromise the independence of potential outcomes and randomly assigned *intended* treatments. The IV framework provides a set of simple strategies to convert comparisons using intended random assignment, i.e., ITT effects, into consistent estimates of the causal effect of treatments delivered.

The easiest way to see how IV solves the compliance problem is in the context of a model with constant treatment effects, i.e., $Y_{1i} - Y_{0i} = \alpha$, for some constant, α. Also, let $Y_{0i} = \beta + \epsilon_i$, where $\beta \equiv E[Y_{0i}]$. The potential outcomes model can now be written

$$Y_i = \beta + \alpha D_i + \epsilon_i, \tag{1}$$

where α is the treatment effect of interest. Note that because D_i is a dummy variable, the regression of Y_i on D_i is just the difference in mean outcomes by delivered treatment status. As noted above, this differences does not consistently estimate α because Y_{0i} and D_i are correlated (equivalently, ϵ_i and D_i are correlated).

The random assignment of intended treatment status, which I'll call Z_i, provides the key to untangling causal effects in the face of treatment dilution and migration. By virtue of random assignment, and the assumption that assigned treatments have no direct effect on potential outcomes other than through delivered treatments, Y_{0i} and Z_i are independent. It therefore follows that

$$E[\epsilon_i | Z_i] = 0, \tag{2}$$

though ϵ_i is not similarly independent of D_i. Taking conditional expectations of (1) with Z_i switched off and on, we obtain a simple formula for the treatment effect of interest:

$$\{E[Y_i | Z_i=1] - E[Y_i | Z_i=0]\}/\{E[D_i | Z_i=1] - E[D_i | Z_i=0]\} = \alpha. \tag{3}$$

Thus, the causal effect of *delivered* treatments is given by the causal effect of *assigned* treatments (the ITT effect) divided by $E[D_i | Z_i=1] - E[D_i | Z_i=0]$.

The denominator in (3) is the difference in compliance rates by assignment status. In the MDVE,

$$E[D_i| Z_i=1]=P[D_i=1| Z_i=1] = .77,$$

that is, a little over three-fourths of those assigned to be coddled were coddled. On the other hand, almost no one assigned to be arrested was coddled:

$$E[D_i| Z_i=0]=P[D_i=1| Z_i=0] = .01.$$

Hence, the denominator of (3) is estimated to be about .76. The sample analog of equation (3) is called a Wald estimator, since this formula first appeared in a paper by Wald (1940) on errors-in-variables problems. The law of large numbers, which says that sample means converge in probability to population means, ensures that the Wald estimator of α is consistent (i.e., converges in probability to α).[5]

The constant-effects assumption is clearly unrealistic. We'd like to allow for the fact that some men change their behavior in response to coddling, while others are affected little or not at all. There is also important heterogeneity in treatment delivery. Some suspects would have been coddled with or without the experimental manipulation, while others were coddled only because the police were instructed to treat them this way. The MDVE is informative about causal effects only on this latter group.

Imbens and Angrist (1994) showed that in a world of heterogeneous treatment effects, IV methods capture the average causal effect of delivered treatments on the subset of treated men whose delivered treatment status can be changed by the random assignment of intended treatment status. The men in this group are called *compliers*, a term introduced in the IV context by Angrist, Imbens, and Rubin (1996). In a randomized drug trial, for example, compliers are those who "take their medicine" when randomly assigned to do so, but not otherwise. In the MDVE, compliers were coddled when randomly assigned to be coddled but would not have been coddled otherwise.

---

[5]Although Wald and other IV estimators are consistent, they are not unbiased. See Angrist and Krueger (2001) for more on the distinction between consistency and unbiasedness in the IV context.

The average causal effect for compliers is called a local average treatment effect (LATE). Formal description of LATE requires one more bit of notation. Define potential treatment assignments D_{0i} and D_{1i} to be individual *i*'s treatment status when Z_i equals 0 or 1. Note that one of D_{0i} or D_{1i} is necessarily counterfactual since observed treatment status is

$$D_i = D_{0i} + Z_i(D_{1i} - D_{0i}). \quad (4)$$

In this setup, the key assumptions supporting causal inference are: (i) conditional independence, i.e., that the joint distribution of $\{Y_{1i}, Y_{0i}, D_{1i}, D_{0i}\}$ is independent of Z_i; and, (ii) monotonicity, which requires that either $D_{1i} \geq D_{0i}$ for all *i* or vice versa. Assume without loss of generality that monotonicity holds with $D_{1i} \geq D_{0i}$. Monotonicity requires that, while the instrument might have no effect on some individuals, all of those affected are affected in the same way. Monotonicity in the MDVE amounts to assuming that random assignment to be coddled can only make coddling more likely, an assumption that seems plausible. Given these two *identifying assumptions*, the Wald estimator consistently estimates LATE, which is written formally as $E[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}]$.[6]

Compliers are those with $D_{1i} > D_{0i}$, i.e., they have $D_{1i}=1$ and $D_{0i}=0$. The monotonicity assumption partitions the world of experimental subjects into three groups: compliers who are affected by random assignment and two unaffected groups. The first unaffected group consists of always-takers, i.e., subjects with $D_{1i}=D_{0i}=1$. The second unaffected group consists of never-takers, i.e., subjects with $D_{1i}=D_{0i}=0$. Because the treatment status of always-takers and never-takers is invariant to random assignment, IV estimates are uninformative about treatment effects for subjects in these groups.

In general, LATE is not the same as ATET, the average causal effect on all treated individuals. Note from equation (4) that the treated can be divided into two groups: the set of subjects with $D_{0i}=1$, and

---

[6]In econometrics, a parameter is said to be "identified" when it can be constructed from the joint distribution of observed random variables. Assumptions that allow a parameter to be identified are called

the set of subjects with $D_{0i}=0$, $D_{1i}=1$, and $Z_i=1$. Subjects in the first set, with $D_{0i}=1$, are always-takers since $D_{0i}=1$ implies $D_{1i}=1$ by monotonicity. The second set consists of compliers with $Z_i=1$. By virtue of the random assignment of Z_i, the average causal effect on compliers with $Z_i=1$ is the same as the average causal effects for all compliers. In general, therefore, ATET differs from LATE because it is a weighted average of two effects: those on always-takers as well as those on compliers.

An important special case when LATE equals ATET is when D_{0i} equals zero for everybody, i.e., there are no always-takers. This occurs in randomized trials with one-sided non-compliance, a scenario that typically arises because no one in the control group receives treatment. If no one in the control group receives treatment, then by definition there can be no always-takers. Hence, all treated subjects must be compliers. The MDVE is (approximately) this sort of experiment. Since we have defined treatment as coddling, and (almost) no one in the group assigned to be arrested was coddled, there are (almost) no always-takers. LATE is therefore ATET, the effect of coddling on the population coddled.[7]

The language of 2SLS

Applied economists typically discuss IV using the language of two-stage least (2SLS), a generalized IV estimator introduced by Theil (1953) in the context of simultaneous equation models. In models without covariates, 2SLS using a dummy instrument is the same as the Wald estimator. In models with exogenous covariates, 2SLS provides a simple and easily-implemented generalization that also allows for multiple instruments and multiple treatments.

Suppose the setup is the same as before, with the modification that we'd like to control for a vector of covariates, X_i. In particular, suppose that if D_i had been randomly assigned as intended, we'd be

---

"identifying assumptions." The identifying assumptions for IV, independence and monotonicity, allow us to construct LATE from the joint distribution of $\{Y_i, D_i, Z_i\}$.

interested in a regression-adjusted treatment effect computed by ordinary least squares (OLS) estimation of

$$Y_i = X_i'\beta + \alpha D_i + \epsilon_i. \quad (5)$$

In 2SLS language, (5) is the structural equation of interest.

The two most likely rationales for including covariates in a structural equation are (i) that treatment was randomly assigned conditional on these covariates, and, (ii) a possible efficiency gain. In the MDVE, for example, the coddling treatment might have been randomly assigned with higher probability to suspects with no prior history of assault. We'd then need to control for assault history in the IV analysis. The efficiency rationale is a consequence of the fact that regression standard errors – whether 2SLS or OLS – are proportional to the variance of the residual, ϵ_i. The residual variance is typically reduced by the covariates, as long as the covariates have some power to predict outcomes.[8]

In principle, we can construct 2SLS estimates in two steps, each involving an OLS regression. In the *first stage*, the "endogenous" right-hand side variable (treatment delivered in the MDVE) is regressed on the "exogenous" covariates plus the instrument (or instruments). This regression can be written

$$D_i = X_i'\pi_0 + \pi_1 Z_i + \eta_i. \quad (6)$$

The coefficient on the instrument in this equation, π_1, is called the "first-stage effect" of the instrument. Note that the first-stage equation must include exactly the same exogenous covariates as appear in the structural equation. The size of the first-stage effect is a major determinant of the statistical precision of

---

[7]The fact that a randomized trial with one-sided non-compliance can be used to estimate the effect of treatment on the treated was first noted by Bloom (1984).

[8]The causal (LATE) interpretation of IVestimates is similar in models with and without covariates. See Angrist and Imbens (1995) or Abadie (2003) for details.

IV estimates. Moreover, in a model with dummy endogenous variables like the treatment dummy analyzed here, the first-stage effect measures the proportion of the population that are compliers.[9]

In the second stage, fitted values from the first-stage are plugged directly into the structural equation in place of the endogenous regressor. Although the term 2SLS arises from the fact that 2SLS estimates can be constructed from two OLS regressions, we don't usually compute them this way. This is because the resulting standard errors are incorrect. Best practice therefore is to use a packaged 2SLS routine such as may be found in SAS or Stata.[10]

In addition to the first-stage, an important auxiliary equation that is often discussed in the context of 2SLS is the *reduced form*. The reduced form for Y_i is the regression obtained by substituting the first-stage into the structural model for Y_i. In the MDVE, we can write the reduced form as

$$\begin{aligned} Y_i &= X_i'\beta + \alpha[X_i'\pi_0 + \pi_1 Z_i + \eta_i] + \epsilon_i \\ &= X_i'\delta_0 + \delta_1 Z_i + \nu_i. \end{aligned} \tag{7}$$

The coefficient δ_1 is said to be the 'reduced-form effect" of the instrument. Like the first stage, the reduced form is consistently estimated by OLS.

Note that with a single endogenous variable and a single instrument, the causal effect of D_i in the structural model is the ratio of reduced-form to first-stage effects:

$$\alpha = \delta_1/\pi_1.$$

In a randomized trial with imperfect compliance, the reduced-form effect is the ITT effect. More generally, 2SLS second-stage estimates can be understood as a re-scaling of the reduced form. It can also be shown that the significance levels for the reduced-form and the second-stage are asymptotically the

---

[9]Formally, this is because without covariates, $E[D_{1i}-D_{0i}]=\pi_1$. With covariates, $E[D_{1i}-D_{0i}|\ X_i]=\pi_1$ if the first-stage is linear and additive in covariates, and, more generally, $E\{E[D_{1i}-D_{0i}|\ X_i]\}\approx\pi_1$.

[10]See Wooldridge (2003) or another econometrics text for details.

same under the null hypothesis of no treatment effect. Hence, the workingman's IV motto: "If you can't see it in the reduced form, it ain't there."

2SLS Estimates for MDVE with one endogenous variable

The first-stage effect of being assigned to the coddling treatment is .79 in a model without covariates and .77 in a model that controls for a few covariates.[11] These first-stage effects can be seen in the first two columns of Table 2, which report estimates of equation (6) for the MDVE. The reduced form effects of random assignment to the coddling treatment, reported in columns 3 and 4, are about .11, and significantly different from zero with standard errors of .041-.047. The first-stage and reduced form estimates change little when covariates are added to the model, as expected since Z_i was randomly assigned.[12]

The 2SLS estimates associated with these first stage and reduced form estimates are .14-.145. The 2SLS estimates, reported in columns 3-4 of Table 3, are about double the size of the corresponding OLS estimates of the effects of delivered treatments, reported in columns 1-2 of the same table. The OLS estimates are almost certainly too low, probably because delivered treatments were contaminated by selection bias. The reduced form effect of coddling is also too small, relative to the causal effect of coddling *per se*, because non-compliance dilutes ITT effects. As noted above, the 2SLS estimates in this case capture the causal effect of coddling on the coddled, undiluted by non-compliance and unaffected by

---

[11]The covariates are dummies for the presence of a weapon and whether the suspect was under chemical influence, year and quarter dummies for time of follow-up, and dummies for suspects' race (nonwhite and mixed).

[12]For simplicity, I discuss these estimates as if they were constructed in the usual way, i.e., by estimating equations (5), (6), and (7) using micro data. In reality, I was unable to locate or construct the original variable on re-offense outcomes in the MDVE public-use data sets (Berk and Sherman, 1993). I therefore generated my own micro-data on re-offense from the logit coefficients reported in Berk and Sherman (1988). By construction, my data set has the same joint distributions of $\{Y_i, D_i\}$, and $\{Y_i, Z_i\}$ as the original data. The observations on $\{D_i, Z_i, X_i\}$ are taken directly from the original data. First-stage estimates are therefore unaffected by the use of artificial data on Y_i.

selection bias.[13] The 2SLS estimates point a dramatic increase in re-offense rates due to coddling (the mean re-offense rate was .18). The magnitude of this effect is clearly understated by alternative estimation strategies.

2SLS estimates with two endogenous variables

The analysis so far looks at the MDVE as if it involved a single treatment. I now turn to a 2SLS model that more realistically allows for distinct causal effects for the two types of coddling that were randomly assigned, separation and advice. A natural generalization of equation (5) incorporating distinct causal effects for these two interventions is

$$Y_i = X_i'\beta + \alpha_a D_{ai} + \alpha_s D_{si} + \epsilon_i, \tag{8}$$

where D_{ai} and D_{si} are dummies that indicate delivery of advice and separation. As before, because of the endogeneity of delivered treatments, OLS estimates of equation (8) are likely to be misleading.

Equation (8) is a structural model with two endogenous regressors, D_{ai} and D_{si}. We also have two possible instruments, Z_{ai} and Z_{si}, dummy variables indicating random assignment to advice and delivery as intended treatments. The corresponding first-stage equations are

$$D_{ai} = X_i'\pi_{0a} + \pi_{aa}Z_{ai} + \pi_{as}Z_{si} + \eta_{ai} \tag{9a}$$

$$D_{si} = X_i'\pi_{0s} + \pi_{sa}Z_{ai} + \pi_{ss}Z_{si} + \eta_{si}, \tag{9b}$$

For models without covariates, the second-stage estimates using my data should be identical to the corresponding microdata estimates. For models with covariates, the estimates using my data should be similar.

[13]It bears emphasizing that even though treatments and outcomes are dummy variables, I used linear models throughout (as opposed to nonlinear models like logit or probit). A simple justification for the use of linear probability models is that marginal effects generated by nonlinear models are likely to be indistinguishable from OLS regression coefficients. More generally the use of 2SLS to estimate linear probability models with dummy endogenous variables is justified by the fact that linear 2SLS estimates have a robust causal interpretation that is insensitive to the possible nonlinearity induced by dummy dependent variables. For example, the interpretation of IV as estimating LATE is unaffected by the fact that the outcome is a dummy. Likewise, consistency of 2SLS estimates is unaffected by the possible nonlinearity of the first-stage conditional expectation function, $E[D_i| X_i, Z_i]$. For details, see Angrist (2001).

where π_{aa} and π_{as} are the first-stage effects of the two instruments on delivered advice, D_{ai}, and π_{sa} and π_{ss} are the first-stage effects of the two instruments on delivered separation, D_{si}.

The reduced form equation for this two-endogenous-variables setup is obtained by substituting (9a) and (9b) into equation (8). Similarly, the second stage is obtained by substituting fitted values from the first-stages into the structural equation.[14] Note that in a model with two endogenous variables we must have at least two instruments for the second stage estimates to exist.[15] Assuming the second stage estimates exist, which is equivalent to saying that the structural equation is identified, the 2SLS estimates in this case can be interpreted as capturing the covariate-adjusted causal effects of each delivered treatment on those who comply with random assignment.

Random assignment to receive advice increased the likelihood of actually receiving this treatment by .78. Assignment to the separation treatment also increased the likelihood of receiving advice, but this effect is small and not significantly different from zero. These results can be seen in columns 1-2 of Table 4, which report the estimates of first-stage effects from equation (9a). The corresponding estimates of equation (9b), reported in columns 3-4 of the table, show that assignment to the separation treatment increased delivered separation rates by about .72, while assignment to advice had almost no effect on the likelihood of receiving the separation treatment. The reduced form effects of random assignment to receive advice range from .088-.097, while the reduced form estimates of random assignment to be separated are about .13. The reduced form estimates are reported in columns 5-6 of the table.

---

[14]With multiple endogenous variables, the second stage estimates can no longer be obtained as the ratio of reduced form to first-stage coefficients, but rather solve a matrix equation. Again, the best strategy for real empirical work is to use packaged 2SLS software.

[15]The second stage has a regression design matrix with number of columns equal to $\dim(X_i)+2$. This matrix must be of full column rank for the second stage to exist. The rank of the design matrix is equal to the number of linearly independent columns in the matrix. This can be no more than $\dim(X_i)$ plus the number of instruments, since the fitted values used in the second step are linear combinations of X_i and the instruments. Hence the need for at least K instruments when there are K endogenous variables.

OLS and 2SLS estimates of the two-endogenous-variables model are reported in Table 5. Interestingly, the OLS estimates of the effect of delivered advice on re-offense rates are small and not significantly different from zero. The OLS estimates of the effect of being separated are more than twice as large and significant. Both of these results are reported in columns 1-2 of the table. In contrast with the OLS effects, the 2SLS estimates of the effects of both types of treatment are substantial and at least marginally significant. For example, the 2SLS estimate of the impact of the advice intervention is .107 (s.e.=.059) in a model with covariates. The 2SLS estimate of the impact of separation is even larger, at around .17.

As in the model with a single endogenous variable, the reduced-form estimates of intended treatment effects are larger than the corresponding OLS estimates of delivered treatment effects, and the 2SLS estimates are larger than the corresponding reduced forms. The gap between OLS and 2SLS is especially large for the advice effects, suggesting that the OLS estimates of the effect of receiving advice are more highly contaminated by selection bias than the OLS estimates of the effect of separation. Moreover, the difference between the separation and advice treatment effects is much larger when estimated by 2SLS than in the reduced form.

Models with variable treatment intensity and observational studies

IV methods are not limited to the estimation of the effects of binary, on-or-off treatments. Many experimental evaluations are concerned with the effects of interventions with variable treatment intensity, i.e., the effects of an endogenous variable that takes on ordered integer values. IV analyses of such interventions include Krueger's (1999) analysis of experimental estimates of the effects of class size, the Permutt and Hebel (1989) study of an experiment to reduce the number of cigarettes smoked by pregnant women, and the Powers and Swinton (1984) randomized study of the effect of hours of preparation for the GRE.

The studies mentioned above use 2SLS or related IV methods to analyze data from randomized trials where the treatment of interest takes on values like 0, 1, 2, . . . (cigarettes, hours of study) or 15, 16, 17 . . . (class size). Although these papers interpret IV estimates using traditional constant-effects models, the 2SLS estimates they report also have a more general LATE interpretation. In particular, 2SLS estimates of models with variable treatment intensity give the average causal response for compliers along the length of the underlying causal response function. See Angrist and Imbens (1995) for details.

The IV framework also goes beyond randomized trials and can be used to exploit quasi-experimental variation in observational studies. An example from my own work is Angrist (1990) , which uses the draft lottery numbers that were randomly assigned in the early 1970s as instrumental variables for the effect of Vietnam-era veteran status on post-service earnings. Draft lottery numbers are highly correlated with veteran status among men born in the early 1950s, and probably unrelated to earnings for any other reason.

A second example from my portfolio illustrates the fact that instrumental variables need not be randomly assigned to be useful.[16] Angrist and Lavy (1999) used something called Maimonides' Rule to construct instrumental variables for the effects of class size on test scores. The instrument in this case is the class size predicted using Maimonides rule, a mathematical formula derived from the practice in Israeli elementary schools of dividing grade cohorts by integer multiples of 40, the maximum class size (the same rule proposed by Maimonides in his *Mishneh Torah* biblical commentary). This study can be seen as an application of Campbell's (1969) celebrated *regression-discontinuity design* for quasi-experimental research, but also as a type of IV. The extension of IV methods to quasi-experimental criminological research designs seems an especially promising avenue for further work.

---

[16]An illustration of this point from criminology is Levitt's (1997) study of the effects of extra policing using municipal election cycles to create instruments for numbers of police. See also McCrary (2002), who discusses a technical problem with Levitt's original analysis.

REFERENCES

Abadie, Alberto (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2):231-263.

Angrist, Joshua D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80(3):313-335.

Angrist, Joshua D. (1990), "Estimation of Limited dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics* 19 (1), 2-16.

Angrist, Joshua D. and Guido W. Imbens (1995), "Two-Stage Least Squares Estimates of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430):431-442.

Angrist, Joshua D. and Alan B. Krueger (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics, Volume IIIA*, Orley Ashenfelter and David Card, eds. Amsterdam: North-Holland, 1277-1366.

Angrist, Joshua D. and Alan B. Krueger (2001), "Instrumental Variables and the Search for Identification," *Journal of Economic Perspectives*, 15(4):69-86.

Angrist, Joshua D. and Victor C. Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement," *Quarterly Journal of Economics*, 114(2):533-575.

Angrist, Joshua D. and Victor C. Lavy (2002), "The Effect of High School Matriculation Awards - Evidence from Randomized Trials," NBER Working Paper 9389, December.Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434):444-55.

Berk, Richard A. and Lawrence W. Sherman (1993), SPECIFIC DETERRENT EFFECTS OF ARREST FOR DOMESTIC ASSAULT: MINNEAPOLIS, 1981-1982 [Computer file]. Conducted by the Police Foundation. 2nd ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor].

Berk, Richard A. and Lawrence W. Sherman (1988),"Police Response to Family Violence Incidents: An Analysis of an Experimental Design With Incomplete Randomization," *Journal of the American Statistical Association*, 83(401):70-76.

Berk, Richard A., Gordon K. Smyth, and Lawrence W. Sherman (1988), "When Random Assignment Fails: Some Lessons From the Minneapolis Spouse Abuse Experiment," *Journal of Quantitative Criminology*, 4(3):209-223.

Bloom, Howard S. (1984), "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 8(2)225-246.

Boruch, Robert, Dorothy De Moya, and Brooke Snyder (2002), "The Importance of Randomized Field Trials in Education and Related Areas," in *Evidence Matters: Randomized Trials in Education Research*, Frederick Mosteller and Robert Boruch, eds. Washington, D.C.: Brookings Institution Press.

Campbell, Donald T. (1969), "Reforms as Experiments," *American Psychologist*, 24:409-429.

Cook, Thomas D. (2001), "Sciencephobia: Why Education Researchers Reject Randomized Experiments," *Education Next* (www.educationnext.org), Fall, 63-68.

Efron, Bradley and D. Feldman (1991), "Compliance as an Explanatory Variable in Clinical Trials," *Journal of the American Statistical Association*, 86(413):9-17.

Farrington, David P. (1983), "Randomized Experiments on Crime and Justice," in *Crime and Justice*, Michael H. Tonry and Norval Morris, eds. Chicago: University of Chicago Press.

Farrington, David P. and Brandon C. Welsh (2005), "Randomized Experiments in Criminology: What Have We Learned in the Last Two Decades?" *Journal of Experimental Criminology*, 1:9-38.

Gartin, Patrick R. (1995), “Dealing with Design Failures in Randomized Field Experiments: Analytic Issues Regarding the Evaluation of Treatment Effects,” *Journal of Research in Crime and Delinquency*, 32(4):425-445.

Holland, Paul W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81(396):945-970.

Imbens, Guido W. and Joshua D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2):467-475.

Krueger, Alan B. (1999), “Experimental Estimates of Education Production Functions,” *Quarterly Journal of Economics*, 114(2):497-532.

Levitt, Steven D. (1997), “Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime,” *American Economic Review*, 87(3):270-290.

McCrary, Justin (2002), “Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime: Comment,” *American Economic Review*, 92(4):1236-1243.

Permutt, Thomas and J. Richard Hebel (1989), "Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight," *Biometrics*, 45(2):619-622.

Powers, Donald E. and Spencer S. Swinton (1984), "Effects of Self-Study for Coachable Test Item Types," *Journal of Educational Psychology*, 76(2):266-278.

Rezmovic, Eva L., Thomas J. Cook, and L. Douglas Dobson (1981), “Beyond Random Assignment: Factors Affecting Evaluation Integrity,” *Evaluation Review*, 5(1):51-67.

Rubin, Donald B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66:688-701.

Rubin, Donald B. (1977), "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2:1-26.

Sherman, Lawrence W. and Berk, Richard A. (1984), “The Specific Deterrent Effects of Arrest for Domestic Assault,” *American Sociological Review*, 49(2):261-272.

Theil, Henri (1953), “Repeated Least Squares Applied to Complete Equation Systems,” The Hauge: Central Planning Bureau.

Wald, Abraham (1940), “The Fitting of Straight Lines If Both Variables Are Subject to Error,” *Annals of Mathematical Statistics*, 11:284-300.

Weisburd, David L. (2003), “Ethical Practice and Evaluation of Interventions in Crime and Justice: the Moral Imperative for Randomized Trials,” *Evaluation Review*, 27(3):336-354.

Weisburd, David L., Cynthia Lum, and Anthony Petrosino (2001), “Does Research Design Affect Study Outcomes in Criminal Justice?” *Annals of the American Academy of Political and Social Science*, 578(6):50-70.

Woodbury, Stephen A. and Robert G. Spiegelman (1987), “Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois,” *American Economic Review*, 77(4):513-530.

Wooldridge, Jeffrey (2003), *Introductory Econometrics: A Modern Approach*, Cinncinati, Ohio: Thomson South-Western.

Table 1: Assigned and Delivered Treatments
in Spousal Assault Cases

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: The table shows statistics from Sherman and Berk (1984), Table 1.

Table 2: First Stage and Reduced Forms for Model 1

Endogenous Variable is Coddled				
	First-Stage		Reduced Form (ITT)	
	(1)	(2)*	(3)	(4)*
Coddled-assigned	0.786 (0.043)	0.773 (0.043)	0.114 (0.047)	0.108 (0.041)
Weapon		-0.064 (0.045)		-0.004 (0.042)
Chem. Influence		-0.088 (0.040)		0.052 (0.038)
Dep. Var. mean	0.567 (coddled-delivered)		0.178 (failed)	

Notes: The table reports OLS estimates of the first-stage and reduced form for Model 1 in the text. *Other covariates include year and quarter dummies, and dummies for non-white and mixed race.

Table 3: OLS and 2SLS Estimates for Model 1

Endogenous Variable is Coddled				
	OLS		IV/2SLS	
	(1)	(2)*	(3)	(4)*
Coddled-delivered	0.087 (0.044)	0.070 (0.038)	0.145 (0.060)	0.140 (0.053)
Weapon		0.010 (0.043)		0.005 (0.043)
Chem. Influence		0.057 (0.039)		0.064 (0.039)

Notes: The Table reports OLS and 2SLS estimates of the structural equation in Model 1. *Other covariates include year and quarter dummies, and dummies for non-white and mixed race.

Table 4: First Stage and Reduced Forms for Model 2

Two Endogenous Variables: Advise, Separate						
	First Stages				Reduced Form (ITT)	
	Advised		Separated			
	(1)	(2)	(3)	(4)	(5)	(6)
Advise-assigned	0.778 (0.039)	0.766 (0.039)	0.035 (0.043)	0.035 (0.043)	0.097 (0.054)	0.088 (0.046)
Separate-assigned	0.044 (0.038)	0.031 (0.039)	0.717 (0.042)	0.715 (0.043)	0.130 (0.053)	0.127 (0.046)
Weapon		-0.038 (0.036)		-0.031 (0.039)		-0.001 (0.042)
Chem. Influence		-0.068 (0.032)		-0.018 (0.035)		0.051 (0.038)
Dep. Var. Mean	0.283 (adv.-deliver)		0.283 (sep.-deliver)		0.178 (failed)	

Notes: The table reports OLS estimates of the first-stage and reduced form for Model 2 in the text. In addition to the covariates reported in the table, these models include year and quarter dummies, and dummies for non-white and mixed race.

Table 5: OLS and 2SLS Estimates for Model 2

Two Endogenous Variables: Advise, Separate				
	OLS		IV/2SLS	
	(1)	(2)	(3)	(4)
Advise-assigned	0.047 (0.052)	0.019 (0.046)	0.116 (0.068)	0.107 (0.059)
Separate-assigned	0.126 (0.052)	0.120 (0.046)	0.174 (0.073)	0.174 (0.063)
Weapon		0.015 (0.043)		0.008 (0.043)
Chem. Influence		0.052 (0.039)		0.061 (0.039)
Test: Advise=Separate	F=1.87 p=.170	F=4.14 p=.043	F=.64 p=.420	F=1.14 p=.290

Notes: The Table reports OLS and 2SLS estimates of the structural equation in Model 2. In addition to the covariates reported in the table, these models include year and quarter dummies, and dummies for non-white and mixed race.