

# Instrumental Variables Estimation in Political Science: A Readers' Guide \*

Allison J. Sovey and Donald P. Green  
Yale University

September 18, 2009

## Abstract

The use of instrumental variables regression in political science has evolved from an obscure technique in the 1970s to a staple of the political science tool kit. Yet the surge of interest in the instrumental variables method has led to implementation of uneven quality. After providing a brief overview of the method and the assumptions on which it rests, we chart the ways in which these assumptions are invoked in practice in political science. We review more than one hundred articles published in the *APSR*, the *AJPS*, and *WP* over a twenty-four year span, categorizing applications according to the way in which instrumental variables are identified. We discuss in detail two noteworthy applications of instrumental variables regression, calling attention to the statistical assumptions that each invokes. The concluding section proposes reporting standards for instrumental variables results and provides a checklist for readers to consider as they evaluate applications of this method.

---

\*Paper prepared for the 26th Annual Society for Political Methodology Summer Conference, Yale University, July 23-25, 2009. We are grateful to Thad Dunning, Daniel Butler, Jake Bowers, Jan Box-Steffensmeier, and John Bullock for helpful comments. We also thank Mario Chacon and Peter Aronow, who assisted us in data collection and provided valuable suggestions. This project was funded by support from Yale's Institution for Social and Policy Studies. We are responsible for any errors. Comments on this draft are welcome and should be directed to [allison.ovey@yale.edu](mailto:allison.ovey@yale.edu).

# 1 Introduction

Political scientists frequently seek to gauge the effects of independent variables that are measured with error or are systematically related to unobserved causes of the dependent variable. Recognizing that ordinary least squares regression performs poorly in these situations, an increasing number of political scientists since the 1970s have turned to instrumental variables (IV) regression. IV regression in effect replaces the problematic independent variable with a proxy variable that is uncontaminated by error or unobserved factors that affect the outcome. Instrumental variables regression is designed to relax some of the rigid assumptions of OLS regression, but IV introduces assumptions of its own. Whether IV is in fact an improvement over OLS depends on the tenability of those assumptions in specific applications (Bartels, 1991).

In order to judge the tenability of IV assumptions, readers of political science research must have access to two types of information. The first is empirical: readers must be presented with certain basic statistics that shed light on the susceptibility of the IV estimator to bias. The second type of information is theoretical: readers must be presented with a description of the causal parameter to be estimated and an argument explaining why the proposed instrumental variable satisfies the underlying assumptions for consistent estimation. The purpose of this essay is to call attention to the fact that although IV applications in political science have grown more numerous and sophisticated, published applications of IV regression commonly fail to present arguments and evidence that readers need in order to understand and evaluate the statistical conclusions.

We begin by providing a brief overview of the assumptions underlying the use of instrumental variables. We discuss the method first in terms of traditional econometric models and then present some of the more subtle assumptions that are made apparent in models of potential outcomes (Angrist, Imbens and Rubin, 1996). We then examine the ways in which these assumptions are invoked in practice in political science applications, based on a review of more than one hundred articles published in the *American Political Science Review*, the *American Journal of Political Science*, and *World Politics* over a twenty-four year span. We then discuss in detail two noteworthy applications of instrumental variables regression, calling attention to the assumptions that each invokes. The concluding section proposes a set of standards to guide the presentation of instrumental variables estimation

results and a checklist for readers to consider as they evaluate applications of this method.

## 2 Brief Overview of Instrumental Variables Estimation

Instrumental variables estimation is traditionally explicated using structural econometric models (Bowden and Turkington, 1984; Goldberger, 2007), with more recent textbooks using potential outcomes notation as well (Morgan and Winship, 2007; Wooldridge, 2001). The former has the virtue of simplicity, and so we start with it. The latter has the advantage of calling attention to several assumptions that are often implicit or ignored by traditional treatments. This section provides a succinct overview of the logic underlying IV regression; for more detailed statistical exposition, see Murray (2006 *a,b*) and Angrist and Pischke (2008).

The traditional structural equation model posits a linear and additive relationship between a dependent variable ( $Y_i$ ), an endogenous regressor ( $X_i$ ), a set of covariates ( $Q_{1i}, Q_{2i}, \dots, Q_{Ki}$ ), and an unobserved disturbance term ( $u_i$ ). In this model

$$Y_i = \beta_0 + \beta_1 X_i + \lambda_1 Q_{1i} + \lambda_2 Q_{2i} + \dots + \lambda_K Q_{Ki} + u_i \quad (1)$$

the parameter of interest is  $\beta_1$ , the causal effect of  $X_i$  on  $Y_i$ . The effects of the covariates in the model are of secondary importance. Note that this model represents the causal effect ( $\beta_1$ ) as a constant, implying that each observation is equally influenced by the treatment. We will later relax this condition.

A model of this form allows for consistent estimation of ( $\beta_1$ ) via ordinary least squares (OLS) if  $plim \frac{1}{N} \sum X_i u_i = 0$ . The motivation for instrumental variables estimation is that this requirement is violated when  $X_i$  is systematically related to unobserved causes of  $Y_i$ . Violations of this sort commonly occur when factors that predict outcomes are omitted from the regression model or when independent variables are measured with error. Note that one need not believe that Y is causing X in order to have good reason to use IV. Two-way causation is not the only concern.

The instrumental variables estimator is premised on a two-equation model in which the endogenous regressor ( $X_i$ ) is written as a linear function of an

instrumental variable ( $Z_i$ ) and the covariates.

$$X_i = \gamma_0 + \gamma_1 Z_i + \delta_1 Q_{1i} + \delta_2 Q_{2i} + \dots + \delta_K Q_{Ki} + e_i \quad (2)$$

Note that  $Z_i$  is an “excluded” instrumental variable in the sense that it appears in equation (2) but not equation (1)<sup>1</sup>.

The instrumental variables estimator in this case is  $(Z'X)^{-1}Z'Y$  where  $Z$  is a matrix with  $N$  rows and  $K+2$  columns (a vector of ones for the constant, a vector of  $Z_i$  and the  $K$  columns of the  $Q_{Ki}$ ) and  $X$  is a matrix with  $N$  rows and  $K+2$  columns (a vector of ones for the constant, a vector of  $X_i$  and the  $K$  columns of the  $Q_{Ki}$ ). The  $N \times 1$  vector  $Y$  is simply  $Y_i$ . This estimator provides consistent estimates of  $\beta_1$  when three conditions are met.

- $plim \frac{1}{N} \sum Z_i u_i = 0$ .
- $plim \frac{1}{N} Z'X = \Sigma_{XZ}$ , a finite nonsingular<sup>2</sup> matrix.
- $plim \frac{1}{N} Z'Z = \Sigma_{ZZ}$ , a positive definite matrix.

The first condition establishes the validity of the exclusion restriction. Because it involves the relationship between an observed variable ( $Z_i$ ) and an unobserved variable ( $u_i$ ), this key assumption is either justified by a theoretical stipulation about the nature of the instrumental variable or a procedure such as random assignment of the  $Z_i$ , which implies that  $Z_i$  and  $u_i$  will be statistically independent. The second condition permits  $Z'X$  to be inverted in order to form the IV estimator; substantively, this condition means that the instrumental variable ( $Z_i$ ) to some degree predicts the endogenous independent variable ( $X_i$ ). The third condition simply ensures that the instrumental variable ( $Z_i$ ) retains some amount of variance as  $N \rightarrow \infty$ . This condition rules out, for example, a study in which a treatment is assigned to some subjects initially, but all remaining subjects are placed into a control group, causing the variance of  $Z_i$  to converge to zero as  $N \rightarrow \infty$ .

---

<sup>1</sup>In principle, several variables could be used as excluded instrumental variables, in which case two-stage least squares provides more efficient estimates than instrumental variables regression, and the availability of excess instruments allows the researcher to conduct goodness-of-fit tests. Ordinarily, instrumental variables are scarce, and so we focus on the case in which just one excluded instrumental variable enables us to estimate the effect of an endogenous regressor.

<sup>2</sup>Nonsingular means that the matrix has an inverse, which in turn implies that none of the covariates or instruments are exact linear combinations of one another. In effect, we are assuming the absence of perfect collinearity.

When evaluating applications of IV, the first of these assumptions merits special attention. In the context of experimental studies using a so-called encouragement design, subjects may be randomly assigned ( $Z_i$ ) to receive a treatment ( $X_i$ ). Well-known examples of this type of design are randomly assigned encouragements of patients to get a flu vaccination (Hirano et al., 2000) or randomly assigned attempts by canvassers to mobilize voters on the eve of an election (Gerber and Green, 2000). If random assignment transmits its influence on the outcome solely through the mediating variable  $X_i$ , then the first condition is also satisfied. In the case of the flu vaccine study, one could imagine a violation of this condition were it the case that encouragement to get a vaccine, rather than the vaccine itself, affected health outcomes. In the case of voter mobilization experiments, this assumption would be violated, for example, if another mobilization campaign learned of the experimental groups and directed its canvassers to contact the experimenter’s control group.<sup>3</sup>

In non-experimental research, the validity of  $plim \frac{1}{N} \sum Z_i u_i = 0$  is often unclear or controversial. As Dunning (2008a, 288) points out, instrumental variables may be classified along a spectrum ranging from randomly generated to “plausibly random” to “less plausibly random.” In the category of plausibly random are near-random IVs that are determined by forces that have little apparent connection to unmeasured causes of  $Y_i$ . Economists in recent years have generated a remarkable array of these kinds of studies. McCleary and Barro (2006) use distance from the equator as an instrument by which to identify the effect of per capita GDP on religiosity. Duflo and Pande (2007) use land gradient as an instrument for dam construction in explaining poverty. Acemoglu, Johnson and Robinson (2001) use the mortality of colonial settlers to estimate the effect of current institutional arrangements on economic performance. Kern and Hainmueller (Forthcoming) use whether an individual lives near Dresden as an instrument to determine the effect of West German television on political attitudes in East Germany. Whether these instruments qualify as “plausibly random” is a matter of opinion, but at least the authors advance reasoned arguments about why such instruments are independent of unobserved factors that affect the dependent variable. Less plausibly random IVs include variables such as demographic attributes in studies of political attitudes or higher-order powers of the predictors in

---

<sup>3</sup>Violations of this assumption may also occur due to sample attrition. See Barnard et al. (2003)

equation (1). These variables are dubbed instruments as a matter of stipulation, often without any accompanying argumentation. Whether such variables are truly unrelated to the unmeasured causes of  $Y_i$  is uncertain and perhaps even doubtful.

Even well-reasoned IV specifications may involve modeling uncertainty, and this modeling uncertainty should be reflected in the standard errors associated with IV estimates. However, it is difficult to quantify this uncertainty, and current reporting conventions essentially ignore it.<sup>4</sup> In effect, the estimated standard errors presuppose no modeling uncertainty at all. Thus, it is left to the reader of instrumental variables regression to form an opinion about the plausibility of the exclusion restrictions and to adjust the reported standard errors accordingly.

Ideally, such opinions are guided by authors' explanations for why the exclusion restrictions are plausible. Unfortunately, as documented below, explanations of this sort are frequently absent from political science publications using IV regression. It should be noted that the plausibility of the exclusion restriction hinges on theoretical argumentation; it cannot be established empirically. Occasionally, one observes political scientists arguing that an instrument is valid because it does not significantly predict  $Y$  in a regression of  $Y$  on  $X$ ,  $Z$ , and covariates. This is a misguided regression that does not provide reliable information about whether  $Z$  is excludable.<sup>5</sup>

Unlike the question of whether instrumental variables are exogenous, which is theoretical, the second assumption hinges on an empirical relationship between  $Z_i$  and  $X_i$ . If the partial correlation between  $Z_i$  and  $X_i$  (controlling for  $Q$ ) is low, the so-called weak instruments problem can lead to substantial finite sample bias when there is only a slight correlation between  $X_i$  and  $u_i$ . Wooldridge (2002), provides a useful heuristic discussion of the weak instruments problem in the simple case where equation (1) excludes covariates (i.e., all  $\gamma_k = 0$ ). He notes that in this case the probability limit of the IV estimator may be expressed  $plim \hat{\beta}_1 = \beta_1 + \frac{r_{Z_i, u_i} \sigma_{u_i}}{r_{Z_i, X_i} \sigma_{X_i}}$ , where  $r_{AB}$  denotes the correlation between the variables  $A$  and  $B$ . This formula makes clear that although the correlation between the instrumental variable ( $Z_i$ ) and the disturbance term ( $u_i$ ) may be very slight in a given application, the amount of bias may be very large if the correlation between ( $Z_i$ ) and ( $X_i$ ) is

---

<sup>4</sup>For Bayesian methods of quantifying uncertainty, see Gill (2002) and Gerber, Green and Kaplan (2004).

<sup>5</sup>This regression will not yield unbiased estimates of  $Z$ 's effects when  $X$  is endogenous.

also very small. Fortunately, the problem of weak instruments is relatively easy to diagnose. Stock and Watson (2007) suggests conducting an F-test in which the sum of squared residuals from equation (2) is compared to the SSR from a restricted regression that excludes the instrumental variable(s). For a single instrumental variable, F statistics under 10 are thought to suggest a problem of weak instruments.<sup>6</sup>

To this point, we have considered a system of linear equations in which the effect of  $X_i$  is assumed to be constant across all observations. This assumption may fail to hold in a variety of applications. For example, suppose an interest group randomly assigns voters to receive calls designed to persuade them to vote for a particular candidate. It may be that targeted voters who are easy to reach by phone are more responsive to campaign appeals than voters who are hard to reach. Indeed, the campaign may target a particular group precisely because they are both easy to reach and especially responsive to the message. The problem is that IV regression estimates the so-called local average treatment effect (LATE), that is, the average treatment effect among those who would be contacted if assigned to the treatment group but not contacted if assigned to the control group. This local average treatment effect may be different from the average effect in the entire population of voters.

In order to highlight the assumptions that come into play when we allow for heterogeneous treatment effects, we apply the potential outcomes framework presented by Angrist, Imbens and Rubin (1996) to the application described by Albertson and Lawrence (2009) in their study of the effects of viewing a Fox News Special on voters' support for a ballot proposition on affirmative action. In their study, which we revisit in more detail below, subjects who were randomly assigned to the treatment group were encouraged to view the program, and the outcome measure of interest is whether, in the context of a follow-up survey, subjects reported supporting the ballot measure. For ease of exposition, we assume that assignment, treatment, and outcomes are each binary variables. We characterize the dependent variable,  $y_i$ , as a pair of potential outcomes for subject  $i$ :  $y_{i1}$  denotes the subject's

---

<sup>6</sup>In this case, Stock and Watson suggests using limited information maximum likelihood, which, according to Monte Carlo simulations, is less prone to bias and has more reliable standard errors. Another suggestion is to focus on the reduced form regression of  $Y$  on  $Q$  and  $Z$  (Chernozhukov and Hansen, 2008). A permutation methods approach to inference in the presence of weak instruments is presented by Imbens and Rosenbaum (2005).

voting behavior if exposed to the Fox News Special, and  $y_{i0}$  denotes the subject’s response if not exposed to this show. Thus, when classified according to their potential responses to the treatment, there are four possible types of subjects: those who oppose Proposition 209 regardless of whether they are treated or not ( $y_{i1} = 0, y_{i0} = 0$ ), those who support Proposition 209 if treated and not otherwise ( $y_{i1} = 1, y_{i0} = 0$ ), those who oppose Proposition 209 if treated and support it otherwise ( $y_{i1} = 0, y_{i0} = 1$ ), and those who support Proposition 209 regardless of whether they are treated ( $y_{i1} = 1, y_{i0} = 1$ ). Note that we will assume that a person’s response is solely a function of whether they personally are treated; assignments or treatments applied to others have no effect. This is known as the Stable Unit Treatment Value Assumption, or SUTVA. We further assume what Angrist and Pischke (2008, 153) call the independence assumption: the potential outcomes  $y_i$  are independent of assigned treatment. In other words, potential support for the proposition has no bearing on the experimental assignment. In addition to independence, we assume that apart from increasing the probability of viewing, assignment to the treatment group has no effect on the outcome. This is simply a restatement of the exclusion restriction.

We further distinguish among four potential responses to the experimental encouragement to view the show. Using Angrist, Imbens and Rubin (1996)’s terminology, we call “Compliers” those who view the Fox News Special if and only if they are assigned to the treatment group. Those who watch the Special program regardless of whether they are assigned to the treatment group are called “Always-Takers.” Those who do not watch regardless of the experimental group to which they are assigned are called “Never-Takers.” Finally, those who watch only if they are assigned to the control group are called “Defiers.”

[TABLE 1 ABOUT HERE]

Based on this setup, there are sixteen possible combinations of  $y_i$  and  $x_i$ , which is to say sixteen possible kinds of subjects. Table 1 describes each of the possible voter types. Each type comprises a share  $\pi_j$  of the total subject population, with  $\sum_k^{16} \pi_j = 1$ . When we speak of the Complier Average Causal Effect, we refer to the causal effect of viewing the Fox News Special among those who are Compliers. From Table 1, we see that the Complier Average Causal Effect is

$$E[y_{i1} - y_{i0} | i \in \text{Compliers}] = \frac{\pi_6 - \pi_7}{\pi_5 + \pi_6 + \pi_7 + \pi_8} \quad (3)$$

The denominator of this equation represents the proportion of Compliers. Without ample numbers of Compliers, the experimenter faces the equivalent of a weak instruments problem: random encouragement to view the Fox News Special will be weakly correlated with actual viewing.

Empirically, we are limited by the fact that we do not observe  $y_{i1}$  and  $y_{i0}$  for the same individuals. Instead, one outcome is observed, and the other remains counterfactual. In order to estimate the Complier Average Causal Effect (CACE), a researcher may conduct a randomized experiment. Suppose that the researcher randomly assigns subjects to the treatment group ( $Z_i = 1$ ) or the control group ( $Z_i = 0$ ). Among those assigned to the treatment group, some watch the Fox News Special ( $Z_i = 1, X_i = 1$ ) and others do not ( $Z_i = 1, X_i = 0$ ). Among those assigned to the control group, some watch the Fox News Special ( $Z_i = 0, X_i = 1$ ) and others do not ( $Z_i = 0, X_i = 0$ ).

A randomized experiment provides estimates of several potentially useful quantities. We will observe the average outcome among those assigned to the treatment group, the average outcome among those assigned to the control group, and the proportion of each experimental group that is actually treated. As Angrist, Imbens and Rubin (1996) point out, even this information is insufficient to identify the causal effect without further assumptions. In particular, we assume that the population contains no Defiers (i.e.,  $\pi_{13} = \pi_{14} = \pi_{15} = \pi_{16} = 0$ ).<sup>7</sup> This stipulation is known as the monotonicity assumption (Angrist, Imbens and Rubin, 1996): the probability of viewing the Fox News Special given that one is assigned to the treatment group is greater than or equal to the probability of viewing given that one is assigned to the control group.

With these assumptions in place, the researcher observes the rate of Proposition 209 support in the assigned treatment group ( $Z_i = 1$ ) and in the assigned control group ( $Z_i = 0$ ). As the number of control group observations  $N_c \rightarrow \infty$  the observed rate of support in the assigned control group ( $\hat{V}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} y_{i0}$ ) may be expressed as

$$plim_{N_c \rightarrow \infty} \hat{V}_c = \pi_3 + \pi_4 + \pi_7 + \pi_8 + \pi_{10} + \pi_{12} \quad (4)$$

because we have assumed that there are no Defiers. Similarly, as the number of observations increases, the support rate in the assigned treatment group

---

<sup>7</sup>Strictly speaking, the monotonicity assumption also holds if there are only Defiers. In essence, we must assume the existence of Compliers or Defiers but not both.

converges in probability to

$$plim_{N_t \rightarrow \infty} \hat{V}_t = \pi_3 + \pi_4 + \pi_6 + \pi_8 + \pi_{10} + \pi_{12} \quad (5)$$

The fraction of the population who watches the Fox News debate if and only if encouraged to do so ( $\alpha$ ) is estimated in a consistent manner by the proportion of the assigned treatment group who watches minus the proportion of the assigned control group who watches:

$$plim_{N \rightarrow \infty} \hat{\alpha} = plim_{N_t \rightarrow \infty} (\hat{\pi}_5 + \hat{\pi}_6 + \hat{\pi}_7 + \hat{\pi}_8 + \hat{\pi}_9 + \hat{\pi}_{10} + \hat{\pi}_{12}) - plim_{N_c \rightarrow \infty} (\hat{\pi}_9 + \hat{\pi}_{10} + \hat{\pi}_{11} + \hat{\pi}_{12}) = \pi_5 + \pi_6 + \pi_7 + \pi_8 \quad (6)$$

Combining equations 4, 5, and 6, the estimator

$$plim_{N \rightarrow \infty} \frac{\hat{V}_t - \hat{V}_c}{\hat{\alpha}} = \frac{\pi_6 - \pi_7}{\pi_5 + \pi_6 + \pi_7 + \pi_8} \quad (7)$$

provides a consistent estimate of the Complier Average Causal Effect defined above. Another way to summarize this result is to say that assuming (1) SUTVA, (2) the exclusion restriction, (3) a nonzero effect of encouragement on actual treatment, and (4) monotonicity, IV regression provides a consistent estimate of the Complier Average Causal Effect. One can also characterize this treatment effect as the “local average treatment effect” in the sense that it refers to the effect only among those who are induced to view the TV special by the experimental encouragement (Angrist, Imbens and Rubin, 1996).

Note that the researcher will not know the identities of the Compliers. In the treatment group, Compliers look just like Always-Takers, and in the control group Compliers look just like Never-Takers. Moreover, Compliers’ share of the population depends on the nature of the experimental encouragement. If the encouragement is weak, there may be relatively few Compliers. The broader point is that different experiments may obtain different results depending on who is induced to comply with the encouragement. This point is glossed over in traditional presentations of instrumental variables estimation, which assume constant treatment effects. Once heterogeneous treatment effects are admitted as a possibility, caution must be exercised when extrapolating from an estimated LATE to other settings or populations.<sup>8</sup>

---

<sup>8</sup>In defense of traditional presentations, some analysts do consider heterogeneous effects (Moffitt, 1996).

### 3 IV in the Political Science Literature

In political science, the quantity and quality of instrumental variables applications have evolved considerably over time. In this section, we describe trends in the use of instrumental variables in leading political science journals. We analyze articles appearing in the *American Political Science Review*, the *American Journal of Political Science* and *World Politics* during the period 1985-2008. The *APSR* and *AJPS* were chosen because articles in these journals employ instrumental variables methods more often than do articles in other political science journals listed in JSTOR during the period of interest. Additionally, we included *World Politics* to ensure that our sample was representative of literature in international relations and international political economy. Articles spanning the years 1985-2007 in the *AJPS*, 1985-2005 in the *APSR* and 1985-2003 in *World Politics* were obtained through searches in Jstor. For more recent articles, the journals were searched directly. In the case of *World Politics*, the Project Muse website was searched. Search terms included “instrumental variable,” “instrumental variables,” “2sls,” “3sls,” and “stage least squares.”<sup>9</sup> A detailed listing of the articles retrieved in this search may be found in a supplementary appendix. Table 2 presents summary statistics of the 102 articles for which instrumental variables methods were mentioned in the body of the text. The articles are divided into four chronological groups: 1985-1990, 1991-1996, 1997-2002, and 2003-2008. Each article is further classified according to three criteria: the way in which exclusion restrictions are justified, whether the model is just-identified or over-identified, and whether first stage results are presented.

Our content analysis classified authors’ justifications for the choice of instruments into one of the following categories: “Experiment,” “Natural Experiment,” “Theory,” “Lag,” “Empirics,” “Reference” or “None.” The “Experiment” category comprises instrumental variables that are formed through random assignment. In principle, instruments that are formed by random assignment satisfy Assumption 1, although any given application may suffer from problems that undermine random assignment, such as sample attrition that affects the treatment group more severely than the control group.

The next category, “Natural Experiment,” includes instruments that are not formed using random assignment, but can still be considered “plausibly

---

<sup>9</sup>Other searches were tried such as “IV,” “endogenous” and “instrument” but these yielded far too many unrelated results to be useful.

random.” Note that this category only includes one article, as only Lassen (2004) employed a near-random intervention as an instrumental variable. Lassen exploits a Copenhagen referendum on decentralization which was carried out in four of fifteen city districts in order to estimate the effect of information on voter turnout. The districts were created for the purpose of the experiment and four districts, chosen such as to be representative of the city, introduced local administration for a four year period. The instrument used counts as “plausibly random” since it was created using near-random assignment.

The third category, “Theory,” includes articles in which authors provide a theoretical explanation for the validity of their exclusion restrictions. In other words, the author presented some type of reasoned argument for why the chosen instrument should be uncorrelated with the error term. An example of a theoretical argument which falls in this category is Tsai (2007), which uses the existence of rural Chinese temple activity before 1949 to instrument for the current existence of a temple manager to explain public goods provision. Tsai argues that “because of the nearly complete eradication of community temples and collective temple activities and the radical social upheaval during the Maoist period it is unlikely that a history of pre-communist temple activity has influenced the current performance of village governments in any way except by making the current existence of temple groups more likely by providing a familiar template for newly organizing social groups” (Tsai, 2007, 366). Each article in this category contains justifications such as Tsai’s; however, the strength of argumentation about the validity of the exclusion restrictions varies widely. For example, many authors use variables such as age, gender, or education as instruments, arguing that these should be unrelated to the error term in their regression equation. Our content analysis took a permissive view of what constitutes a persuasive theoretical justification.

The fourth category, “Lag”, includes IVs that were generated by lagging the dependent or independent variables. In many cases, one can make compelling theoretical arguments for using a lagged variable as an instrument. For example, Gerber (1998) presents a model estimating the effect of campaign spending on Senate election outcomes. To estimate incumbent vote percentage, the endogeneity of campaign spending must be dealt with. He instruments for campaign spending using lagged spending by incumbents and challengers, arguing that “due to the staggered nature of Senate elections, the previous race and the current race rarely involve the same incumbent

or challenger. The variable is therefore free from the criticism that might be applied to lagged spending by the same candidate, namely, that specific candidate attributes are correlated with both the regression error and past fundraising levels” (Gerber, 1998, 405). Again, instrumental variables in this category must be viewed with caution as their validity depends on the strength of the author’s argumentation.

The fifth “Empirics” category includes IVs that were selected based on the results of an empirical test. Such tests often include regressions of Y on X and Z to show no correlation between Y and Z, or regressing X on Z to determine the most highly correlated instruments. Such empirical tests do not represent valid methods for establishing the validity of the exclusion restrictions. The first regression is biased insofar as X is endogenous (suspicions about endogeneity are presumably what impelled the researcher to turn to IV regression); the second regression says nothing about whether Z is independent of the disturbance term.

Our sixth category, “Reference,” contains articles in which the author explains the validity of his exclusion restrictions by citing another author’s work. For example, Lau and Pomper (2002) select the same instruments as Gerber (1998) and therefore merely cite Gerber’s work rather than providing a full justification for their selection.

Finally, the category “None” includes all articles where no justification for the exclusion restrictions is provided. Two coders evaluated each article in order to confirm the lack of explanation.

[TABLE 2 ABOUT HERE]

Table 2 displays some encouraging trends. Firstly, it is clear that the percentage of articles that provide some justification for the choice of instruments has risen substantially. Articles falling under the “Experiment,” “Natural Experiment,” “Theory,” “Lag,” and “Reference” categories have all risen overtime. Collectively, the articles in these categories have increased from a low of 14% between 1991-1996 to 56% in the most recent period. Specifically, we can note that the use of experiments and natural experiments gained prominence, from 0% in early periods to 6% most recently.<sup>10</sup> Another

---

<sup>10</sup>We also analyze the data by running a multinomial logistic regression of categories on time. This analysis yields substantively similar results, showing a marked shift in articles falling under the “theory” category over time.

encouraging sign is the rising percentage of just-identified models. Apparently, the realization that valid instruments are hard to find and defend is gradually causing political scientists to become more discriminating in their choice of instruments. These numbers clearly point to the trend of increasing sophistication among political scientists in selection and implementation of instrumental variables methods. Reporting practices have also evolved over time. The percentage of articles reporting first stage results increases from a low of 7% between 1991-1996 to 33% between 2003-2008. In absolute terms, there is still much room for improvement, and only a fraction of those who report first-stage results assess statistically whether instruments are weak or whether overidentifying restrictions are satisfied. Nevertheless, trends in argumentation and presentation show increasing attention to aspects of IV regression that affect the persuasiveness of the results. We now turn to two noteworthy examples of especially creative uses of IV. The fact that both sets of authors have made their replication data available means that their use of IV can be evaluated in depth.

## **4 A Closer Look at Examples of IV Applications**

In this section, we carefully examine two illustrative applications. The first uses random assignment as an instrumental variable and illustrates the special considerations that arise with noncompliance and attrition. The second uses a near-random intervention, change in rainfall, as an instrumental variable and illustrates the special considerations that arise when applying IV in a setting where assumptions of ignorability and SUTVA may be violated.

### **4.1 Application 1: IV Regression and a Randomized Experiment with Noncompliance**

Those who study the effects of media exposure outside the laboratory confront the problem of selective exposure: people decide whether to watch a TV program, and there may be important unmeasured differences between viewers and non-viewers. In an innovative attempt to address the selection problem, Albertson and Lawrence (2009) analyzed a randomized experiment in which survey respondents were randomly encouraged to view a Fox News

debate on affirmative action on the eve of the 1996 presidential election. Shortly after the election, these respondents were re-interviewed. The post-election questionnaire asked respondents whether they viewed the Fox News debate and whether they supported Proposition 209, which dealt with affirmative action. The authors report that 45.2% of the 259 people who were re-interviewed in the treatment group watched the half-hour program, as compared to 4.4% of the 248 respondents who were re-interviewed in the control group. The F-statistic implied by this first-stage regression is 142.2, which allays any concerns about weak instruments.

Albertson and Lawrence appropriately modeled the relationship between media exposure and support for Proposition 209 in a manner that does not presuppose that exposure is exogenous. Their two equation system is

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{8}$$

$$X_i = \gamma_0 + \gamma_1 Z_i + e_i \tag{9}$$

where  $Y_i$  is support for Proposition 209,  $X_i$  is exposure to the Fox News debate, and  $Z_i$  is the random assignment to the treatment group.<sup>11</sup> Using  $Z_i$  as an instrumental variable, Albertson and Lawrence’s IV regression results show a substantively strong but statistically insignificant relationship between program viewing and support for the proposition. Viewers were 8.1 percentage-points more likely to support the ballot measure, with a standard error of 9.3 percentage-points.<sup>12</sup>

Several features of this study are noteworthy from the standpoint of statistical inference. First, the estimand is the local average treatment effect of viewing the program on those who were induced to do so by the interviewers’ encouragement, which included a follow-up letter containing \$2 and a reminder to watch in the form of a refrigerator magnet. It seems reasonable to suppose that these blandishments only increased respondents’ probability of viewing the debate, which implies that we can safely assume monotonicity (i.e., no Defiers). It follows that Compliers constitute 40.8% of this sample.

Second, the ignorability restriction stipulates that the treatment and control groups are identical except for the effects of the program. In defense of

---

<sup>11</sup>The authors also include an array of covariates, but we exclude these for ease of exposition.

<sup>12</sup>In other analyses, the authors find that viewing the program had no effects on voter turnout or on attitude polarization, although there is some evidence that viewers felt more informed about the issue.

this assumption, one would argue that random assignment created groups that, in expectation, have identical potential outcomes. In addition, it seems reasonable to suppose that the follow up letter and accompanying payment had no direct effect on support for Proposition 209. On the other hand, the exclusion restriction is potentially threatened by attrition from the treatment and control groups. We do not know whether rates of attrition are similar in the two experimental groups or, more generally, whether the causes of attrition are similar. If attrition operates differently in the two groups and if attrition is related to support for Proposition 209, the IV estimates may be biased.

To investigate whether attrition presents a problem for their research design, we use Albertson and Lawrence’s replication data to conduct a randomization check. Their data set only contains information for those who completed both the pre-test and the post-test, and the question is whether attrition introduced noticeably imbalance among pre-treatment covariates. A regression of treatment assignment on the demographic variables used in their study does not yield any significant predictors of treatment assignment. (The demographic variables in our regression include: Party Identification, Interest in Politics, Watch National News, Read Newspapers, Education, Income, Gender, White, and dummy variables for missing values of control variables). The F-statistic,  $F(16,490)=1.24$ ,  $p=.23$ , implies that attrition is not related to pre-treatment observables.

A third concern involves the measurement of compliance. Respondents self-report whether they viewed the Fox News debate, and the difference between the treatment and control group viewing rates forms the denominator of the IV estimator. A potential concern is that those in the treatment group may over-report whether they viewed the program in order to appear to comply with interviewers’ encouragement. This form of measurement error will cause researchers to overstate the proportion of Compliers and therefore to underestimate the local average treatment effect. As the authors note, future applications of this encouragement design may wish to insert some specific recall measures to gauge the reliability of these self-reports.

## **4.2 Application 2: IV Regression and a Natural Experiment**

Miguel, Satyanath and Sergenti (2004) present a natural experiment that has

attracted a great deal of attention in political science due to the clever way they address the identification problem. The authors use variation in rainfall growth (percentage change in rainfall from the previous year) to instrument for economic growth in order to estimate the impact of economic conditions on civil conflict. This approach attempts to overcome the problems of correlation between economic growth and unobserved causes of conflict, which has plagued other observational studies. The first stage of their model estimates the relationship between rainfall and income growth:

$$growth_{it} = \alpha_{1i} + X'_{it}b_1 + c_{1,0}\Delta R_{it} + c_{1,1}\Delta R_{i,t-1} + \delta_{1i}year_t + e_{1it} \quad (10)$$

The authors focus on the incidence of civil war in country  $i$  in year  $t$  ( $c_{it}$ ) using the PRIO/Uppsala database, but also present results using the onset of conflict. Current and lagged rainfall growth ( $\Delta R_{i,t}$  and  $\Delta R_{i,t-1}$ ), is used to instrument for per capita economic growth ( $growth_{it}$ ), controlling for other country characteristics ( $X_{it}$ ). Country fixed effects ( $a_i$ ) and country-specific time trends are also included in most specifications. Miguel et al find a strong positive relationship between rainfall and income growth in all of the specifications they present.

The second-stage equation estimates the impact of income growth on the incidence of violence:

$$conflict_{it} = \alpha_{2i} + X'_{it}\beta_2 + \gamma_{2,0}growth_{it} + \gamma_{2,1}growth_{i,t-1} + \delta_{2i}year_t + \epsilon_{2it} \quad (11)$$

They perform both IV-2SLS estimation and a nonlinear two-stage procedure and find that economic growth significantly reduces the likelihood of civil conflict.

It is instructive to review the assumptions on which this claim rests. First, consider the estimand. Unless one is prepared to assume that effects of a one-unit change in economic growth are the same regardless of how economic growth comes about, the instrumental variables estimator may be said to gauge the local average treatment effect of rainfall-induced growth. In his critique of Miguel, Satyanath and Sergenti (2004), Dunning (2008a) argues that growth in different economic sectors may have different effects on conflict and that rainfall helps illuminate the growth-induced effects of the agricultural sector. Relaxing the assumption of homogeneous treatment effects forces more cautious extrapolations from the results. The results tell us not about the effects of economic growth but of a particular type of economic growth.

[TABLE 3 ABOUT HERE]

A second assumption is that rainfall is a near-random source of variation in economic growth. In a natural experiment, “it is assumed that some variable or event satisfies the criterion of ‘randomness,’ the event or variable is orthogonal to the unobservable and unmalleable factors that could affect the outcomes under study” (Rosenzweig and Wolpin, 2000, 827). In this case, the exogeneity of rainfall is uncertain. If variation in rainfall growth were truly random, it should be unpredictable. One cannot know whether a variable is truly random, but one can examine whether this variable’s associations with other observable variables are consistent with random assignment. Using the replication dataset that Miguel et al. provide with their article, we show in Table 3 that factors such as oil reserves, mountainous terrain, and lagged GDP predict rainfall growth. Rainfall could still be exogenous conditional on the covariates in the model. However, the reason using rainfall as an instrument is intuitively appealing is that we think of rainfall as patternless. It appears that rainfall growth is systematically related to a range of other observable variables, and therefore we have to assume we have just the right covariates in order to isolate the random component of rainfall.

A further estimation concern is that the treatment that rainfall induces in one country may have consequences for the economic growth in another country, creating potential SUTVA violations. For example, assignment to drought in one country could make another country’s products more scarce and therefore more valuable. These possible SUTVA violations can produce biased estimates, which, importantly, could be biased in either direction.<sup>13</sup>

## 5 A reader’s checklist

Having reviewed the assumptions underlying instrumental variables regression, both in general and with regard to specific applications, we conclude with a checklist for readers to consider as they evaluate argumentation and evidence.

---

<sup>13</sup>Even if there were no spill-over effects, countries that are geographically proximal are likely to share weather assignments. The fact that rainfall is randomly “assigned” to geographic locations has potentially important consequences for the estimated standard errors. Miguel et al. cluster by country, but this is not the same as clustering by geographic weather patterns and may lead to an underestimation of the standard errors (Arceneaux, 2005).

1. What is the estimand? A basic conceptual question is whether treatment effects are homogenous. Instrumental variables may influence participation in treatment for only a subset of the units, identifying the local average treatment effect. If homogenous treatment effects are assumed, then  $LATE = ATT = ATE$ . However, in the presence of heterogeneous treatment effects, “IVLS asymptotically estimates the impact of the exogenous portion of treatment, not the endogenous piece or a mixture of endogenous and exogenous pieces” (Dunning, 2008*b*, 298). When drawing inferences from IV results, the reader should attend to the manner in which variation in the treatment is induced and whether this type of variation is likely to have idiosyncratic or general effects.

2. Are the exclusion restrictions valid? The reader should pay special attention to potential violations of the assumption that the instrument can have no effect on the outcome except through the treatment. Note that this assumption cannot be verified directly from the data, as it relates quantities that cannot be jointly observed. In applications which use instrumental variables the reader should take note as to whether exclusions are empirically, procedurally or theoretically derived. If theoretically derived, the reader must examine how plausible the exclusion restrictions are. Could the instrument affect the outcome through a channel other than the endogenous regressor?

3. Are the instruments weak? “Weak instruments” are instrumental variables whose incremental contribution to R-squared (over and above the contribution of other covariates) in the first stage equation is so low that the risk of bias is severe. Although the precise criteria by which to evaluate the weakness of instrument is subject to debate, the usual rule of thumb is that a single instrumental variable should have an F-statistic of at least 10 in order to avoid appreciable weak instruments bias. In the case of a single instrumental variable, this criterion means that the first stage t-ratio must be greater than 3.16.

4. Are those assigned to treatment always at least as likely to receive treatment? The assumption of monotonicity states that for all subjects, the probability the subject is treated is at least as great when the subject is in the treatment group as when the subject is in the control group. This assumption rules out the existence of Defiers who are less likely to be treated if assigned to the treatment group. This assumption is satisfied by design in experiments where the treatment is only available to the treatment group. However, in other applications such as natural experiments, this assumption can be violated since the researcher does not assign units to treatment ex

ante. For example, in the natural experiment discussed above, a positive change in rainfall may not always lead to higher economic growth. More rain could actually impede growth for very wet regions. If this were the case, the assumption of monotonicity would be violated, leading to potentially biased results.

5. Are units affected by the treatment of others? Violations of the Stable Unit Treatment Value Assumption, or SUTVA, may occur when treatments vary from one unit to the next. The most common source of concern occurs when outcomes for one unit depend upon whether or not other units receive the treatment. SUTVA violations may lead to biased estimates. The sign and magnitude of the bias depend on the way in which treatment effects spill over across observations.

These items, while important, do not exhaust the list of concerns, and one could easily expand the checklist to include complications arising from limited dependent variables (Maddala, 1985), sample attrition (Manski, 1990), or clustered assignment to treatment (Wooldridge, 2003). But even with a longer list of evaluative criteria, the reader in political science currently confronts a basic challenge: most publications that use instrumental variables regression provide none of the necessary argumentation or evidence by which one might evaluate the statistical claims. If authors could be encouraged to address the abbreviated checklist presented above, the quality of exposition - and one hopes, investigation - would improve substantially.

It should be stressed that the use of instrumental variables regression is likely to grow dramatically in years to come, and with good reason. IV represents a valuable method of addressing problems of selection bias and unobserved heterogeneity. By providing a checklist for readers to consider as they critically evaluate applications, we in no way wish to imply that IV is inferior to other estimation approaches. On the contrary, instrumental variables regression is extraordinarily useful both as an estimation approach and as a framework for research design. The reason to read instrumental variables applications with care is that this type of identification-oriented research deserves special attention.

Table 1: Classification of Target Population in Fox News Study

Group No.	Type	Watches Fox News Special if Assigned to Treatment?	Watches Fox News Special if Assigned to Control?	Supports Prop. 209 if Watches Debate? ( $y_{i1}$ )	Supports Prop 209 if Does Not Watch Debate? ( $y_{i0}$ )	Share of the Population
1	Never-takers	No	No	No	No	$\pi_1$
2	Never-takers	No	No	Yes	No	$\pi_2$
3	Never-takers	No	No	No	Yes	$\pi_3^{ab}$
4	Never-takers	No	No	Yes	Yes	$\pi_4^{ab}$
5	Compliers	Yes	No	No	No	$\pi_5$
6	Compliers	Yes	No	Yes	No	$\pi_6^a$
7	Compliers	Yes	No	No	Yes	$\pi_7^b$
8	Compliers	Yes	No	Yes	Yes	$\pi_8^{ab}$
9	Always-takers	Yes	Yes	No	No	$\pi_9$
10	Always-takers	Yes	Yes	Yes	No	$\pi_{10}^{ab}$
11	Always-takers	Yes	Yes	No	Yes	$\pi_{11}$
12	Always-takers	Yes	Yes	Yes	Yes	$\pi_{12}^{ab}$
13	Defiers	No	Yes	No	No	$\pi_{13}$
14	Defiers	No	Yes	Yes	No	$\pi_{14}^b$
15	Defiers	No	Yes	No	Yes	$\pi_{15}^a$
16	Defiers	No	Yes	Yes	Yes	$\pi_{16}^{ab}$

<sup>a</sup>This share of the population supports Proposition 209 if assigned to the treatment group.

<sup>b</sup>This share of the population supports Proposition 209 if assigned to the control group.

Table 2: Characteristics of IV Applications Over Time

Date Group	1985-1990	1991-1996	1997-2002	2003-2008
Justification %				
Experiment	0	0	4	6
Natural Experiment	0	0	0	3
Theory	9	7	14	31
Lag	9	7	14	11
Reference	5	0	3	5
Empirics	9	0	17	5
None	68	86	48	39
% Just-identified	8	13	24	22
% Report First Stage	23	7	31	33
Number of Articles	22	15	29	36

Table 2 summarizes more than one hundred articles published in the *American Political Science Review*, the *American Journal of Political Science* and *World Politics* over a twenty-four year span, categorizing them according to the way the IVs are identified. Note that the percentages in each date group add with rounding error to 100%  
Explanation of Categories:

Experiment: IVs that were generated through a random assignment process

Natural Experiment: IVs that were generated through a quasi-random assignment process

Theory: Articles in which the author provided a theoretical explanation for the validity her exclusion restrictions

Lag: IVs that were generated by lagging the dependent or independent variable

Empirics: IVs that were selected based on the results of an empirical test (such as regressing Y on X and Z to show no correlation or regressing X on Z to determine the strongest instruments)

Reference: Articles in which the author explains the validity of his exclusion restrictions by citing another author's work

None: No justification provided

Table 3: Assessing whether rainfall growth is predicted by other covariates

Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)
Economic Growth t-1	-.465*** (.127)			-.531*** (.143)		
Economic Growth t-2		-.107 (.104)			-.114 (.120)	
Economic Growth t-3			.354** (.163)			.396** (.174)
Terms of Trade t-1	.035 (.061)			.034 (.064)		
Terms of Trade t-2		-.046 (.065)			-.051 (.068)	
Terms of Trade t-3			.102* (.056)			.087 (.057)
Log (GDP per capita), 1979	.002 (.011)	.002 (.006)	.010 (.007)			
Democracy (Polity IV) t-1	.001 (.001)	.002 (.001)	.001 (.001)			
Ethnolinguistic Fractionalization	-.044 (.054)	.000 (.047)	-.007 (.050)			
Religious Fractionalization	.033 (.071)	.055 (.059)	.054 (.054)			
Oil Exporting Country	.017 (.025)	.007 (.020)	-.002 (.019)			
Log (national population), t-1	-.11 (.009)	.010 (.008)	-.003 (.009)			
Log (mountainous)	.006 (.009)	.013* (.007)	.012 (.007)			
Country fixed effects	No	No	No	Yes	Yes	Yes
Country specific time trends	Yes	Yes	Yes	Yes	Yes	Yes
$R^2$	.039	.043	.053	.053	.052	.071
p-values from F-test of Economic Growth	.0007	.3141	.0361	.0007	.255	.255
p-values from F-test of Terms of Trade	.5624	.4796	.0775	.6033	.151	.151
p-values from F-test of Economic Growth, Terms of Trade	.003	.4830	.0581	.0027	.5231	.0666
Root MSE	.209	.212	.211	.213	.216	.214
Observations	661	661	661	661	661	661

Table 3 uses the Miguel et al. replication dataset to show that factors such as oil reserves, mountainous terrain, and lagged GDP predict rainfall growth.

Note: A country-specific year trend is included in all specifications (coefficients not reported) \* 90% confidence, \*\* 95% confidence, \*\*\* 99% confidence

## References

- Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91(5):1369–1401.
- Albertson, Bethany and Adria Lawrence. 2009. "After the Credits Roll: The Long-Term Effects of Educational Television on Public Knowledge and Attitudes." *American Politics Research* 37(2):275–300.
- Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.
- Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Arceneaux, Kevin T. 2005. "Using Cluster Randomized Field Experiments to Study Voting Behavior." *The Annals of the American Academy of Political and Social Science* 60(1):169–179.
- Barnard, J., C.E. Frangakis, J.L. Hill and D.B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98(462):299–324.
- Bartels, Larry M. 1991. "Instrumental and "Quasi-Instrumental" Variables." *American Journal of Political Science* 35(3):777–800.
- Bowden, Roger J. and Darrell A. Turkington. 1984. *Instrumental variables*. Cambridge University Press, Cambridge; New York.
- Chernozhukov, Victor and Christian Hansen. 2008. "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters* 100(1):68 – 71.
- Duflo, Esther and Rohini Pande. 2007. "Dams." *Quarterly Journal of Economics* 122(2):601–646.
- Dunning, Thad. 2008a. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2):282–293.

- Dunning, Thad. 2008*b*. “Model Specification in Instrumental-Variables Regression.” *Political Analysis* 16(3):290–302.
- Gerber, Alan. 1998. “Estimating the Effect of Campaign Spending on Senate Election Outcomes Using Instrumental Variables.” *The American Political Science Review* 92(2):401–411.
- Gerber, Alan and Donald Green. 2000. “The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment.” *The American Political Science Review* 94(3):653–663.
- Gerber, Alan, Donald Green and Edward Kaplan. 2004. The Illusion of Learning From Observational Research. In *Problems and Methods in the Study of Politics*, ed. Ian Shapiro, Rogers Smith and Tarek Massoud. New York: Cambridge University Press pp. 251–73.
- Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman and Hall.
- Goldberger, Arthur S. 2007. *A Course in Econometrics*. Harvard University Press.
- Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin and Xiao-Hua Zhou. 2000. “Assessing the Effect of an Influenza Vaccine in an Encouragement Design .” *Biostatistics* 1(1):69–88.
- Imbens, Guido W. and Paul R. Rosenbaum. 2005. “Robust, accurate confidence intervals with a weak instrument: quarter of birth and education.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(1):109+.
- Kern, Holger Lutz and Jens Hainmueller. Forthcoming. “Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes.” *Political Analysis* .
- Lassen, David D. 2004. “The Effect of Information on Voter Turnout: Evidence from a Natural Experiment.” *SSRN eLibrary* .
- Lau, Richard R. and Gerald M. Pomper. 2002. “Effectiveness of Negative Campaigning in U.S. Senate Elections.” *American Journal of Political Science* 46(1):47–66.

- Maddala, G. S. 1985. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.
- Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review* 80(2):319–323.
- McCleary, Rachel M. and Robert J. Barro. 2006. "Religion and Economy." *The Journal of Economic Perspectives* 20(2):49–72.
- Miguel, Edward, Shanker Satyanath and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112(4):725–753.
- Moffitt, Robert A. 1996. "Identification of Causal Effects Using Instrumental Variables: Comment." *Journal of the American Statistical Association* 91(434):462–465.
- Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. Cambridge University Press.
- Murray, Michael P. 2006a. "Avoiding Invalid Instruments and Coping with Weak Instruments." *The Journal of Economic Perspectives* 20(4):111–132.
- Murray, Michael P. 2006b. "The Bad, the Weak, and the Ugly: Avoiding the Pitfalls of Instrumental Variables Estimation." *SSRN eLibrary* .
- Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. "Natural "Natural Experiments" in Economics." *Journal of Economic Literature* 38(4):827–874.
- Stock, James H. and Mark W. Watson. 2007. *Introduction to Econometrics: International Edition, 2/E*. Pearson Higher Education.
- Tsai, Lily L. 2007. "Solidary Groups, Informal Accountability, and Local Public Goods Provision in Rural China." *American Political Science Review* 101(02):355–372.
- Wooldridge, Jeffrey. 2002. *Introductory Econometrics: A Modern Approach, 4e*. South-Western College Pub.
- Wooldridge, Jeffrey M. 2001. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *The American Economic Review* 93(2):133–138.