This is a preprint of an article published in [Sovey, Allison J., and Donald P. Green. 2010. Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science* 55(1): 188-200]. Pagination of this preprint may differ slightly from the published version.

Instrumental Variables Estimation in Political Science: A Readers' Guide

Allison J. Sovey Yale University
Donald P. Green Yale University

The use of instrumental variables regression in political science has evolved from an obscure technique 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 100 articles published in the American Journal of Political Science, the American Political Science Review, and World Politics over a 24-year span. 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 and provides a checklist for readers to consider as they evaluate applications of this method.

Political scientists frequently seek to gauge the effects of independent variables that are measured with error or are systematically related to unobserved determinants 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 help readers judge the tenability of IV assumptions, researchers must provide pertinent evidence and argumentation. Readers must have access to certain basic statistics that shed light on the susceptibility of the IV estimator to bias. Readers also need a description of the causal parameter to be estimated and an argument explaining why the proposed instrumental variable satisfies the requirements 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 evidence and arguments that enable readers 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 100 articles published in the *American Political Science Review,* the *American Journal of Political Science,* and *World Politics* over a 24-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 IV estimation results and a checklist for

Allison J. Sovey is a Ph.D. student in Political Science, Yale University, 115 Prospect Street, Rosenkranz Hall, Room 437, New Haven, CT 06520 (allison.sovey@yale.edu). Donald P. Green is A. Whitney Griswold Professor of Political Science, Yale University, 115 Prospect Street, Rosenkranz Hall, Room 437, New Haven, CT 06520 (donald.green@yale.edu).

An earlier version of this article was prepared for the 26th Annual Society for Political Methodology Summer Conference, Yale University. We are grateful to Jake Bowers, John Bullock, Daniel Butler, Thad Dunning, Andrew Gelman, Holger Kern, and Jan Box-Steffensmeier for helpful comments. We also thank Peter Aronow and Mario Chacon, 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.

readers to consider as they evaluate applications of this method.

Brief Overview of Instrumental Variables Estimation

Instrumental variables estimation is traditionally explicated using structural econometric models (Bowden and Turkington 1984; Theil 1971), with more recent textbooks using potential outcomes notation as well (Morgan and Winship 2007; Wooldridge 2002). 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 in or ignored by traditional treatments. This section provides a succinct overview of the logic underlying IV regression; for more detailed statistical exposition, see Murray (2006a), Gelman and Hill (2006), 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 exogenous covariates (Q_{1j}, Q_{2i}.... Q_{Ki}), and an unobserved disturbance term (u_i), each indexed with the subscript i to refer to observations 1 through N. In this model

$$Y_i = \beta_0 + \beta_1 X_i + \lambda_1 Q_{1i} + \lambda_2 Q_{2i} + \cdots + \lambda_k Q_{ki} + u_i \quad (1)$$

the parameter of interest is β_1 ,the causal effect of X_i on Y_i. The effects of the covariates in the model are of secondary importance. Notice that this model represents the causal effect (β_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 via ordinary least squares (OLS) if $plim \frac{1}{N}\Sigma X_i u_i = 0$. In other words, as the sample size approaches infinity, OLS will converge on the true parameter (β_1) so long as the covariance between X_i and u_i approaches zero. 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 related to X_i that predict outcomes are omitted from the regression model or when independent variables are measured with error (Wooldridge 2002, chap. 5). One need not believe that Y_i is causing X_i 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 (Xi) is written as a linear function of an instrumental variable (Zi) and the covariates.1

$$X_i = \gamma_0 + \gamma_1 Z_i + \delta_1 Q_{1i} + \delta_2 Q_{2i} + \cdots + \delta_k Q_{ki} + e_i \quad (2)$$

Here Z_i is an "excluded" instrumental variable in the sense that it appears in equation (2) but not equation (1).[2]

The instrumental variables estimator in this case may be obtained by two-stage least squares: regress X_i on Z_i and the covariates; use the coefficients from this first-stage regression to generate predicted values of X_i; and regress Y_i on the predicted values of X_i as well as the covariates. This estimator presupposes that Z_i is not an exact linear combination of the covariates in the first stage; if it were, the predicted values of X_i would be collinear with the covariates in the second stage, and the estimator would be undefined. Beyond this simple mechanical requirement, instrumental variables regression generates consistent estimates of when two conditions are met. The first is that the covariance between Z_i and u_i goes to zero as N becomes infinite. When critics question the validity of an instrument, they are challenging whether Z_i is truly unrelated to unobserved factors that affect Y_i.

The validity of an instrument may be challenged on various grounds, depending on the research design. In the context of experimental studies using a so-called encouragement design, subjects maybe 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). The fact that encouragements are randomly assigned means that Z_i is independent of other preexisting causes of Y_i, which makes Z_i a potentially valid instrument. For Z_i to be valid, however, it must transmit its influence on the outcome solely through the mediating

[1]The inclusion of exogenous covariates is optional in experimental analysis, as random assignment of the instrumental variable ensures that it is statistically independent of the disturbance regardless of whether covariates are included in the model. In observational studies, the inclusion of covariates usually makes more plausible the assumption that the near-random instrumental variable is independent of the disturbance. One concern, however, is that covariates may not be exogenous, in which case including them may bias the IV estimates.

[2]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 (Wooldridge 2002). 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.

variable X_i. 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.[3] In such cases, Z_i affects Y_i through some channel other than X_i.

In nonexperimental research, the validity of this assumption is often unclear or controversial. As Dunning (2008, 288) points out, instrumental variables may be classified along a spectrum ranging from "plausibly random" to "less plausibly random." In the category of plausibly random are IVs that are determined by forces that have little apparent connection to unmeasured causes of Y_i. Researchers in recent years have generated a remarkable array of these kinds of studies. 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 (2009) 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 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 (Bartels 1991), 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 (Gerber, Green, and Kaplan 2004). Reported standard errors, in other words, 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 argumentation; it cannot be established empirically. Occasionally, one observes political scientists arguing that an instrument is valid because it does not significantly predict Y_i in an OLS regression of Y_i on X_i, Z_i, and covariates. This misguided regression does not provide reliable information about whether Z_i is excludable.[4]

The second assumption is that the covariance of Z_i and X_i (after partialling out the covariance that each variable shares with the covariates) converges to some nonzero quantity as N becomes infinite. Unlike the question of whether instrumental variables are valid, which is largely theoretical, the second assumption can be tested in finite samples based on the 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 even when there is only a slight correlation between X_i and u_i. Wooldridge (2009, 514) 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 as $\beta_1 = \beta_1 + \frac{r_{z_1 u_i}}{r_{z_1 x_i}} \frac{\sigma_{u_i}}{\sigma_{x_i}}$, 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 also very small. Fortunately, the problem of weak instruments is relatively easy to diagnose. Stock and Watson (2007) suggest conducting an F-test that compares the sum of squared residuals from two nested models: equation (2) versus 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.[5]

[3]Violations of this assumption may also occur due to sample attrition. See Barnard et al. (2003) and Manski (1990). We revisit this point below.

[4]This regression will not yield unbiased estimates of Z_i's effects when X_i is endogenous.

[5]In this case, Stock and Watson suggest 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_i on Q_{ki} and Z_i (Chernozhukov and Hansen 2008). A permutation approach to inference in the presence of weak instruments is presented by Imbens and Rosenbaum (2005).

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 discussed 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 experiment, which we discuss 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 as a pair of potential outcomes for subject i: y_{i1} denotes the subject's 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 ($y_{i} = 0$, $y_{i0} = 0$), 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 he or she personally is treated; assignments or treatments applied to others have no effect. This requirement is known as the Stable Unit Treatment Value Assumption, or SUTVA (Rubin 1978). We further assume what Angrist and Pischke (2008, 153) call the independence assumption: the potential outcomes (y_{i0}, y_{i1}) are independent of assigned treatment. 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.

Turning now to potential outcomes associated with receiving the treatment, we further distinguish among four potential responses to the experimental encouragement to view the show. Using Angrist, Imbens, and Rubin's (1996) 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."

Based on this setup, there are 16 possible combinations of y_i and X_i, which is to say 16 possible kinds of subjects. Table 1 describes each of the possible voter types. Each type comprises a share π_j of the total subject population, with $\sum \pi_j = 1$. When we speak of the complier average causal effect (CACE), 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

$$E[y_{i1} - y_{i0} | i \in \text{Compliers}] = \frac{\pi_c}{\pi_c + \pi_d + \pi_n + \pi_a} \quad (3)$$

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, 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

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	π_1
2	Never-takers	No	No	Yes	No	π_2
3	Never-takers	No	No	No	Yes	π_3^{ab}
4	Never-takers	No	No	Yes	Yes	π_4^{ab}
5	Compliers	Yes	No	No	No	π_5
6	Compliers	Yes	No	Yes	No	π_6^{a}
7	Compliers	Yes	No	No	Yes	π_7^{b}
8	Compliers	Yes	No	Yes	Yes	π_8^{ab}
9	Always-takers	Yes	Yes	No	No	π_9
10	Always-takers	Yes	Yes	Yes	Yes	π_{10}^{ab}
11	Always-takers	Yes	Yes	No	Yes	π_{11}
12	Always-takers	Yes	Yes	Yes	Yes	π_{12}^{ab}
13	Defiers	No	Yes	No	No	π_{13}
14	Defiers	No	Yes	Yes	No	π_{14}^{b}
15	Defiers	No	Yes	No	Yes	π_{15}^{a}
16	Defiers	No	Yes	Yes	Yes	π_{16}^{ab}

[a]This share of the population supports Proposition 209 if assigned to the treatment group.
[b]This share of the population supports Proposition 209 if assigned to the control group.

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., $\pi_{13}= \pi_{14}= \pi_{15}= \pi16= 0$). This stipulation is known as the monotonicity assumption (Angrist, Imbens, and Rubin 1996): no one watches the Fox News Special if and only if he or she is assigned to the control group. With these assumptions in place, the experimental design enables the researcher to identify the complier average causal effect, which is also the local average treatment effect.[6]

The mechanics of this identification result become apparent as one traces the groups depicted in Table 1 as a treatment is administered. 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 because we have assumed that there are no Defiers. Similarly, as the number of observations increases, the support rate in the assigned treatment group converges in probability to

[6]Estimation of the CACE or LATE becomes more complicated when one controls or covariates, although in practice results tend to be similar when IV is applied to experimental data. For a discussion of local average response functions, see Abadie (2003).

group ($\hat{V}_C = \frac{1}{N_c}\Sigma_{i=1}^{N_c} y_{i0}$) may be expressed as

$$plim_{Nc \rightarrow \infty} \hat{V}_C = \pi_3 + \pi_4 + \pi_6 + \pi_8 + \pi_{10} + \pi_{12}. \quad (4)$$

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

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

The fraction of the population who watches the Fox News debate if and only if encouraged to do so (α) 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:

$$\begin{aligned} plim_{N \rightarrow \infty} \hat{\alpha} &= (\pi_5 + \pi_6 + \pi_7 + \pi_8 + \pi_9 + \pi_{10} + \pi_{11} + \pi_{12}) \\ &\quad -(\pi_9 + \pi_{10} + \pi_{11} + \pi_{12}) \\ &= \pi_5 + \pi_6 + \pi_7 + \pi_8. \end{aligned} \quad (6)$$

Combining equations (4), (5), and (6), the estimator

$$\text{plim}_{N \rightarrow \infty} \frac{V_t - V_c}{\hat{\alpha}} = \frac{\pi_8 - \pi_7}{\pi_5 + \pi_6 + \pi_7 + \pi_8} \quad (7)$$

provides a consistent estimate of the compiler average causal effect defined above. Another way to summarize this result is to say that assuming (1) the instrument is independent of potential outcomes, (2) the exclusion restriction, (3) a nonzero effect of encouragement on actual treatment, (4) monotonicity, and (5) SUTVA, IV regression provides a consistent estimate of the complier average causal effect. The CACE is also the local average treatment effect, or the average effect among those induced to view the TV special by the experimental encouragement (Angrist, Imbens, and Rubin 1996).

It should be stressed 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 will be relatively few Compliers. The broader point is that even very large experiments may generate different results depending on which sorts of people are induced to comply with the encouragement. This point is glossed over in most 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. Even within a given sample, the average treatment effect among Compliers may not generalize to non-Compliers.

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. We included *World Politics* as well 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."[7] A detailed listing of the articles retrieved in this search may be found in the 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 were divided into four chronological groups: 1985-90, 1991-96, 1997-2002, and 2003-2008. Each article was further classified according to three criteria: the way in which exclusion restrictions are justified, whether the model is just-identified or overidentified, 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 were formed through random assignment, regardless of whether a researcher or government agency conducted the randomization.[8] In principle, instruments that were formed by random assignment satisfy Assumption 1, although any given application may suffer from problems that undermine random assignment, such as sample attrition that afflicts the treatment and control groups differently.

The next category, "Natural Experiment," includes instruments that were not formed using random assignment but can still be considered "plausibly random." It turns out that this category includes just one article, as only Lassen (2004) employed a near-random intervention as an instrumental variable. In order to estimate the effect of information on voter turnout, Lassen exploits a Copenhagen referendum on decentralization that was carried out in four of 15 city districts. The districts were created for the purpose of the experiment, and four districts, chosen to be representative of the city, introduced local administration for a four-year period. The instrument seems "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 authors presented some type of reasoned argument for why the chosen instrument should be uncorrelated with

[7]Other searches were tried, such as "IV," "endogenous," and "instrument," but these yielded far too many unrelated results to be useful.

[8]An example of a government-run experiment, although one that appeared after we completed our content analysis, is Bhavnani's (2009) study of randomly assigned reservations for women candidates in India.

TABLE 2 Characteristics of Published IV Applications Over Time

Type of Justification	1985-1990	1991-1996	1997-2002	2003-2008
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
Total Percent	100%	100%	100%	100%
Number of Articles	22	15	29	36
% Just-identified	8	13	24	22
% Report First Stage	23	7	31	33

Table 2 summarizes more than 100 articles published in the American Political Science Review, the American Journal of Political Science, and World Politics over a 24-year span, categorizing them according to the way the IVs are identified. Percentages within 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 authors provided a theoretical explanation for the validity of their 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 or her exclusion restrictions by citing another author's work.
None: No justification provided.

the error term. An example of a theoretical argument that falls into this category is Tsai (2007), which uses 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 precommunist 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" (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 used 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 theoretical justification.

The fourth category, "Lag," includes IVs that were generated by lagging variables.[9] In certain 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 instrumented 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 category, "Empirics," includes IVs that were selected based on the results of an empirical test. For example, researchers regress Y on X and Z to show no correlation between Y and Z or regress X on Z to determine the most highly correlated instruments. Such empirical tests do not convincingly demonstrate 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

9A lagged variable is a realization of a variable at a previous point in time. For example, lagged campaign spending is the amount spent by a candidate in a previous election.

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 or her exclusion restrictions by citing another author's work. For example, Lau and Pomper (2002) employ the same instruments as Gerber (1998) and 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 displays some encouraging trends. First, it is clear that the percentage of articles that provide some justification for the choice of instruments increased substantially over time. A growing proportion of articles fell into the "Experiment," "Natural Experiment," "Theory," "Lag," and "Reference" categories. Collectively, the articles in these categories increased from a low of 14% between 1991 and 1996 to 56% in the most recent period. During this period, the use of experiments and natural experiments emerged, growing from 0% in early periods to 6% most recently. Another encouraging sign is the rising percentage of just-identified models. Apparently, the realization that valid instruments are hard to find and defend gradually led political scientists to become more discriminating in their choice of instruments.[10] These numbers suggest a trend of increasing sophistication among political scientists in selection and implementation of instrumental variables methods. Reporting practices also became more transparent over time. The percentage of articles reporting the first-stage relationship between Z and X increased from a low of 7% between 1991 and 1996 to 33% between 2003 and 2008.

In absolute terms, however, there is still much room for improvement. Almost half of the articles published as late as 2003-2008 offered no argumentation or deficient argumentation. A minority of articles presented first-stage results, and only a fraction of these assessed statistically whether instruments are weak or whether overidentifying restrictions are satisfied. Nevertheless, there are signs that scholars are becoming more sophisticated in terms of argumentation and presentation. 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.

[10]See Gelman (2009) on alternative identification strategies. Some authors start with a near-random instrument and ask what parameters it might help identify; others start with a parameter they hope to identify and search for a valid instrument.

A Closer Look at Examples of IV Applications

In this section, we closely 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 to non-experimental data.

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 nonviewers. In an innovative attempt to address the selection problem, Albertson and Lawrence (2009) analyzed an 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 reinterviewed. The postelection 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 reinterviewed in the treatment group watched the half-hour program, as compared to 4.4% of the 248 respondents who were reinterviewed in the control group. The F-statistic implied by this firststage regression is 142.2, which allays any concerns about weak instruments.

Albertson and Lawrence appropriately model 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

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad (8)$$

$$X_i = \gamma_0 + \gamma_1 Z_i + e_i \quad (9)$$

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.[11] Using Z_i as an instrumental variable, Albertson and Lawrence's IV regression results

[11] The authors also include an array of covariates, but we exclude these for ease of exposition.

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.[12]

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 among Compliers. Here Compliers are those who potentially watch the program only if assigned to receive an interviewer's 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' propensity to view the debate, which implies that we can safely assume monotonicity (i.e., no Defiers). It follows that Compliers constitute 45.2% - 4.4% = 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 this assumption, one may 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 independence assumption 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 pretest and the posttest, and the question is whether attrition introduced noticeable imbalance among pretreatment 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 nonsignificant F-statistic, $F(16,490) = 1.24$, $p = .23$, is consistent with the null hypothesis that attrition is unrelated to pretreatment 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, researchers using this encouragement design in the future may wish to insert some specific recall measures to gauge the reliability of these self-reports.

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 its clever identification strategy. The authors use variation in rainfall (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 authors focus on the incidence of civil war in country i in year t using the PRIO/Uppsala database that covers 41 countries in 19 years. Current and lagged rainfall growth are used to instrument for per capita economic growth and lagged per capita economic growth controlling for other country characteristics such as religious fractionalization, mountainous terrain, and population. Country fixed effects and country-specific time trends are also included in most specifications. Miguel, Satyanath, and Sergenti find a significant positive relationship between rainfall and GDP growth but acknowledge that change in rainfall falls short of passing the weak instruments test proposed by Stock and Watson (2007, 735) in all of the specifications they present.

The second-stage equation estimates the impact of GDP growth and lagged income growth on the incidence of violence. Their IV/2SLS estimates suggest that current and lagged economic growth significantly reduce the likelihood of civil conflict. This basic pattern of results holds up when the data are analyzed using maximum

[12]In other analyses, the authors find that viewing the program did not affect voter turnout or attitude polarization, although there is some evidence that viewers felt more informed about the issue.

likelihood, suggesting that the weak instruments problem is fairly minor.

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, Dunning (2008) 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 may tell us not about the effects of economic growth but of a particular type of economic growth.

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 can examine whether rainfall's associations with other observable variables are consistent with the hypothesis of random assignment. Using the replication dataset that Miguel, Satyanath, and Sergenti provide with their article, we find that factors such as population, mountainous terrain, and lagged GDP significantly predict rainfall growth or lagged rainfall growth, although these relationships are not particularly strong and the predictors as a group tend to fall short of joint significance.[13] Suppose for the sake of argument that these covariates were found to be systematically related to rainfall growth. Rainfall could still be assumed random 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. If rainfall growth is systematically related to other variables, we have to assume that our regression model includes just the right covariates in order to isolate the random component of rainfall.

A further estimation concern is that rainfall in one country may have consequences for the economic growth in another country, creating potential SUTVA violations. For example, drought in one country could make an other 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.[14]

In sum, these two applications illustrate the kinds of issues that frequently arise in the context of experimental and nonexperimental analysis. When reading experiments that involve noncompliance, one must consider whether the random assignment might influence the outcome for reasons other than the treatment itself. When evaluating experiments more generally, one must be alert to problems such as attrition, which threaten to undermine the comparability of the treatment and control groups. When reading nonexperimental applications, special critical attention must be paid to the assumption that the instrument is unrelated to the disturbance term. Even when the IV is deemed exogenous, the reader should reflect on whether the instrumental variable may transmit its influence on the outcome through causal pathways not specified by the model. Instrumental variables estimation embodies a series of arguments, and the reader must be prepared to critically evaluate these arguments.

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 (summarized in Table 3) for readers to consider as they evaluate argumentation and evidence.

1) What is the estimand? A basic conceptual question is whether treatment effects are homogenous. Instrumental variables regression identifies the local average treatment effect, that is, the average effect among Compliers. If homogenous treatment effects are assumed, then the LATE is the same as the average treatment effect for the sample as a whole. When drawing inferences from IV results, the reader should consider the question of whether results for the Compliers in this particular study are generalizable. For example, do rainfall-induced shocks to economic growth have the same effect on ethnic violence as technology-induced

[13]The significance of these statistical relationships varies depending on model specification. See our supplementary materials for details of the analysis.

[14]Even if there were no spillover 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, Satyanath, and Sergenti cluster by country, but this is not the same as clustering by geographic weather patterns and may underestimate the standard errors (Green and Vavreck 2008).

TABLE 3 Checklist of Issues to AddressWhen Presenting Instrumental Variables Results

Category	Issues to Address	Relevant Evidence and Argumentation
Model	• What is the estimand? • Are the causal effects assumed to be homogenous or heterogeneous?	• Discuss whether other studies using different instruments or populations generate different results.
Independence	• Explain why it is plausible to believe that the instrumental variable is unrelated to unmeasured causes of the dependent variable.	• Conduct a randomization check (e.g., an F-test) to look for unexpected correlations between the instrumental variables and other predetermined covariates. • Look for evidence of differential attrition across treatment and control groups.
Exclusion Restriction	• Explain why it is plausible to believe the instrumental variable has no direct effect on the outcome.	• Inspect the design and consider backdoor paths from the instrumental variable to the dependent variable.
Instrument Strength	• How strongly does the instrument predict the endogenous independent variable after controlling for covariates?	• Check whether the F-test of the excluded instrumental variable is greater than 10. • If not, check whether maximum likelihood estimation generates similar estimates.
Monotonicity	• Explain why it is plausible to believe there are no Defiers, that is, people who take the treatment if and only if they are assigned to the control group.	• Provide a theoretical justification or explain why the research design rules out Defiers (e.g., the treatment is not available to those in the control group).
Stable Unit Treatment Value Assumption (SUTVA)	• Explain why it is plausible to assume that a given observation is unaffected by treatments assigned or received by other units.	• Assess whether there is evidence that treatment effects are transmitted by geographical proximity or proximity within social networks.

growth? The issue of heterogeneous treatment effects is best addressed through replication. Do different instruments generate similar results? Extrapolation becomes increasingly plausible when estimated effects are found to be similar across different groups of Compliers. Replication may be more than one can reasonably expect from a single study, but, where possible, researchers should guide readers' intuitions about heterogeneous effects by describing how other IV approaches have played out in the literature.

2) Is the instrumental variable independent of the potential outcomes? When evaluating the independence assumption, the reader should take note of whether it is justified empirically, procedurally, or theoretically. Empirical justifications that take the form of a statistical test should be read with caution; auxiliary regressions do not provide a direct test of this assumption. One should be especially skeptical when Z_i is proposed as an instrument based on preliminary regressions showing that Z_i has no influence on Y_i controlling for X_i. When authors justify the exclusion restriction based on randomization or a near-random procedure, they should provide some evidence that, consistent with the hypothesis of random assignment, the instrumental variable is weakly predicted by other covariates. If attrition occurs, the researcher should assess whether the loss of treatment and control observations undermines the comparability of these groups. When instrumental variables are proposed on theoretical grounds, readers should reflect on whether the instrument bears a systematic relationship to the disturbance term. For example, McCleary and Barro (2006) use distance from the equator as an instrument by which to identify the effect of per capita GDP on religiosity. Might latitude be correlated with other unmeasured causes of religiosity?

3) Suppose an instrumental variable is deemed exogenous because it is random or near random. Are the exclusion restrictions valid? This assumption implies that the instrument can have no effect on the outcome except through the treatment. In the

case of the Fox News experiment, could it be that opinion change is induced when a person is invited to watch the TV special, regardless of whether he or she in fact watches?

4) 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 an instrument are 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. When instruments fall short of this threshold, researchers are encouraged to check the robustness of their results using other estimators. See Stock and Watson (2007).
5) Does the instrumental variable have a monotonic effect on the treatment? The assumption of monotonicity states that there are no units that receive the treatment if and only if assigned to the control group, ruling out the existence of Defiers. This assumption is satisfied by design in certain experiments where the treatment is only available to the treatment group. However, in other experimental and observational research designs, this assumption is more uncertain. For example, in the Miguel, Satyanath, and Sergenti (2004) study, increased rainfall may not necessarily lead to higher economic growth; more rain could actually impede growth in very wet regions. If so, the assumption of monotonicity would be violated, leading to potentially biased estimates of the local average treatment effect.
6) Are the observations subject to spillover effects? Violations of the Stable Unit Treatment Value Assumption, or SUTVA, occur when outcomes for one unit depend on whether other units receive the treatment. SUTVA violations occur when one observation is affected by another observation's Z_i or X_i. 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 six checklist 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) or clustered assignment to treatment (Wooldridge 2003). But even consulting our abbreviated list of evaluative criteria, the reader in political science currently confronts a basic challenge: most publications that use instrumental variables regression fail to provide the arguments or evidence that readers need in order to evaluate the statistical claims. If authors could be encouraged to consider the abbreviated checklist presented above, the quality of exposition—and, one hopes, estimation—might improve substantially.

The use of instrumental variables regression is likely to grow dramatically in years to come, and with good reason. IV is a valuable method for 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.

References

Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113: 231–63.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economjc Revjew* 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 Assodation* 91: 444–55.

Angrist, Joshua D., and Jorn–Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empïricist's Companion.* Princeton, NJ: Princeton University Press.

Arceneaux, Kevin T. 2005. "Using Cluster Randomized Field Experiments to Study Voting Behavior." *The Annals of the American Academy of Poltical and Social Science* 60(1): 169–79.

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 Assoaation* 98(462): 299–324.

Bartels, Larry M. 1991. "Instrumental and 'Quasi Instrumental' Variables." *American Journal of Poltical Science* 35(3): 777–800.

Bhavnani, Rikhil R. 2009. "Do Electoral Quotas Work after They Are Withdrawn? Evidence from a Natural Experiment in India." *American Political Science Review* 103(1): 23–35.

Bowden, Roger J., and Darrell A. Turkington. 1984. *Instrumental Variables*. Cambridge and New York: Cambridge University Press.

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): 601–46.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2): 282–93.

Gelman, Andrew. 2009. "A Statistician's Perspective on 'Mostly Harmless Econometrics: An Empiricist's Companion,' by Joshua D. Angrist and Jörn–Steffen Pischke." *The Stata Journal* 9(2): 315–20.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Gerber, Alan. 1998. "Estimating the Effect of Campaign Spending on Senate Election Outcomes Using Instrumental Variables." *American Political Science Review* 92(2): 401–11.

Gerber, Alan, and Donald Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3): 653–63.

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, 251–73.

Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster–Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16(2): 138–52.

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. 2009. "Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes." *Political Analysis* 17(4): 377–99.

Lassen, David D. 2004. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment." *American Journal of Political Science* 49(1): 103–18.

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: Cambridge University Press.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80(2): 319–23.

McCleary, Rachel M., and Robert J. Barro. 2006. "Religion and Economy." *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): 725–53.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research).* Cambridge: Cambridge University Press.

Murray, Michael P. 2006. "Avoiding Invalid Instruments and Coping with Weak Instruments." *Journal of Economic Perspectives* 20(4): 111–32.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38(4): 827–74.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1): 34–58.

Stock, James H., and Mark W. Watson. 2007. *Introduction to Econometrics: International Edition.* 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Theil, Henri. 1971. *Principles of Econometrics.* New York: Wiley.

Tsai, Lily L. 2007. "Solidary Groups, Informal Accountability, and Local Public Goods Provision in Rural China." *American Political Science Review* 101(2): 355–72.

Wooldridge, Jeffrey. 2009. *Introductory Econometrics: A Modern Approach.* 4th ed. Mason, OH: South–Western College.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2003. "Cluster–Sample Methods in Applied Econometrics." *American Economic Review* 93(2): 133–38.