Endogeneity Problems with Binary Treatments: A Comparison of Models

Scott J. Basinger
Department of Political Science
447 Philip Guthrie Hoffman Hall
University of Houston
Houston, TX 77204-3011
sjbasinger@uh.edu

Michael J. Ensley
Department of Political Science
302 Bowman Hall
P.O. Box 5190
Kent State University
Kent, OH 44242
mensley@kent.edu

ABSTRACT

This article introduces and compares alternative procedures for controlling endogeneity in a regression model with a binary treatment variable. Results are reported from Monte Carlo simulations designed to test the robustness of these procedures under varying levels of endogeneity and instrument strength. The alternative procedures are further compared by replicating an analysis of the influence of presidential appeals on agency appropriations in the United States. Diagnostics for evaluating endogeneity bias and instrument strength are discussed. Practical advice is offered for researchers who are concerned that a binary treatment variable is endogenous.

1. Introduction

The reality that endogeneity problems pervade the social sciences is both a blessing and a curse. The curse is obvious: uncontrolled endogeneity leads to incorrect estimates of causal effects across widely varying contexts and policies. A prime example is the decades-long debate over the Head Start program's effectiveness, since enrollees may have differed systematically from the population the program sought to aid. The blessing is that endogeneity attracted the attentions of a diverse set of scholars – economists, sociologists, political scientists, etc. – who generated an equally diverse set of solutions. Unfortunately, each solution is accompanied by its own terminology, notation, and canonical applications. Beyond the syntactical baggage, a technique might be tailored to a particular version of endogeneity, undermining consensus on proper methods. In separate contexts, political scientists adapted methods developed by macroeconomists to resolve simultaneity bias in the presence of reciprocal causation (e.g., Hanushek and Jackson 1977; Jackson 2008), and adapted methods developed by policy-, education-, and labor-economists to resolve omitted variables bias (e.g., Achen 1986; Bartels 1991).

Our focus is on a particularly problematic form of endogeneity problem, treatment selection bias,[1] which is encountered when the potentially endogenous variable is binary. Angrist and Pischke (2009:190) refer ominously to linear models for dummy endogenous variables as "forbidden regressions," reflecting the view that the familiar, instrumental-variables practices for dealing with endogeneity must be modified in this context. Thus, the models and diagnostics covered in Jackson's (2008) recent *Handbook* chapter do not apply directly when the endogeneous variable is not continuous. Goldberger, Heckman, and Maddala, among others, conceived of systems-of-equations solutions to this predicament, which will be the primary focus of this paper. However, the treatment selection bias problem also puts a practitioner into contact with the work of Holland, Rosenbaum, and Rubin, who conceived of matched-sampling solutions. Jackson pronounces, "The hope is that these efforts will lead to better tests of causal arguments by avoiding or at least reducing possible endogeneity biases" (2008: 428; see also Sekhon 2008). However,

Ho et al. concede that, "from the point of view of the practical researcher, it looks like a cacophony of conflicting techniques, practices, conventions, and rules of thumb" (2007: 201).

The technical point of this paper, then, is to compare matched-sampling, linear regression, and systems-of-equations methods for studying endogenous treatment effects. In so doing we identify the narrow conditions under which matching methods are sufficient (although not necessarily superior to regression), and we identify conditions under which various methods are to be preferred. The models are engaged in a replication of Canes-Wrone's (2001) analysis of the effect of presidents' public appeals. We conclude by offering guidance on choosing between these various approaches, as well as on combining methods for diagnosing whether endogeneity was a problem in the first place.

2. Overt Bias, Ignorability, and Matched Sampling

Consider the following analogy. A professor has offered to conduct an optional review session prior to a final examination, but he is uncertain whether the session is a valuable use of students' time (not to mention his own time). To evaluate the review session's effectiveness, he asks whether attendance (the treatment, D) improved students' scores on the final exam (the outcome, Y). The prima facie causal effect, also known as the observed average treatment effect (*OATE*), is simply the difference between attendees' and absentees' mean scores:

$$OATE \equiv E(Y \mid D = 1) - E(Y \mid D = 0),$$

where $E(\bullet)$ is the expectations operator. Anyone who has experienced an optional review session, either as a student or as an instructor, should be suspicious about the *a priori* equivalence of the groups of attendees and absentees, due to self-selection. The difference between group averages is likely to produce a biased estimate of the treatment's effectiveness: the bias might be upward if the most intrinsically motivated students are more likely to attend, or downward if students in worst danger of failing are more likely to attend instead. The professor may be misled – although it cannot be known in what direction – because the study is observational (Rosenbaum 2002) or quasi-experimental (Cook and Campbell 1979),

defined as: "a study intended to assess causal effects of treatments where the rule that governs the assignment of treatments to units is at least partially unknown" (Rubin 2006: 7).

Suppose that a well-meaning colleague suggests conducting a controlled experiment, by assigning (or "encouraging") a group of students to attend the review session, and denying that opportunity (or "encouragement") to the remaining students. This procedure adheres to the best advice on circumventing selection problems: "By creating the treatments of interest… [and] by assigning subjects to treatments randomly, the experimenter can be confident (within the limitations established by statistical inference) that any differences observed between subjects assigned to different treatment conditions *must* be caused by differences in the treatments themselves" (Kinder and Palfrey 1993: 11). Unfortunately, forbidding the control group of students from attending the review session may penalize them, akin to denying a treatment to an ill patient. Conversely, requiring the treatment group to attend a review session may harm them by wasting their time or distracting them from other study methods. Lacking the power to exert *ex ante* control over treatment assignment, the quasi-experimental researcher must confront a basic problem of "separating the effect of a treatment from those due to non-comparability between the average units in each treatment group" (Cook and Campbell 1979: 6). To draw a valid inference about the causal effect of a treatment, a practitioner might be required to exert *ex post* control to account for potential biases in treatment selection.

The counterfactual model of causality (Morgan and Winship 2007), often known as the Rubin Causal Model,[2] provides a tidy way to distinguish treatment effects from selection effects in quasi-experimental studies. The model is based on the idea that for each subject, one can conjecture two potential or "what-if" outcomes that might be observed depending on whether the subject is treated. For a subject i, let Y_i denote the observed outcome. Let D_i denote the observed treatment, such that $D_i = 1$ if she receives treatment (attends the review session), and $D_i = 0$ if she does not. Denote her potential or hypothetical outcomes as $Y_{1i} = (Y_i \mid D_i = 1)$ if she is assigned to the treatment group, and $Y_{0i} = (Y_i \mid D_i = 0)$ if she is assigned to the control group. Unfortunately, because the analyst makes a single observation for each individual $[Y_i = (1–D_i)(Y_{0i}) + (D_i)(Y_{1i})]$, individual-level causal effects cannot be calculated – a

predicament that Holland (1986) names "the fundamental problem of causal inference." The implication is that one must compute aggregate causal effects, comparing two or more distinct groups. The practitioner might wish to know the "average treatment effect on the treated":

$$ATET \equiv E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1) = E(Y_1 - Y_0 \mid D = 1).$$

This quantity is the difference between the average outcome that actually attained for the treated, and the average outcome that would have attained had the identical individuals not received treatment. One cannot estimate this quantity directly, because we observe $E(Y_0 \mid D = 0)$ instead of $E(Y_0 \mid D = 1)$. Therefore, *OATE* might differ from *ATET*:

$$E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0) = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1) + \{E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)\}$$

$$OATE = ATET + \{E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)\}.$$

The bracketed term in the above equality is treatment selection bias. *OATE* estimates *ATET* accurately only if $E(Y_0 \mid D = 1) = E(Y_0 \mid D = 0)$, which requires that the *observed* average outcome the control group is no different from the *hypothetical* average outcome that would have attained absent the treatment.

Controlled, random assignment's role in experimental design is to ensure that the assignment of the treatment will be independent of the outcome, so that the bracketed term equals zero *a priori*. If *ex ante* control through random assignment is unethical or impossible or impractical,[3] or even if the practitioner is concerned with efficiency, then the practitioner might exert *ex post* control, using background knowledge to achieve conditional independence of the treatment.[4] To illustrate this point, suppose we denote the average treatment effect by τ, the constant by α, and the disturbance term by ε_i. Under the assumption of treatment independence, we would initially assume ε_i is distributed Normal with zero mean. The observed response will equal:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

Conditional expectations for treatment group and control group, respectively, can be written:

$$E(Y_i \mid D_i = 1) = \alpha + \tau + E(\varepsilon_i \mid D_i = 1)$$

$$E(Y_i \mid D_i = 0) = \alpha + E(\varepsilon_i \mid D_i = 0)$$

The "difference estimator" (Stock and Watson 2007) equals treatment effect plus selection bias:

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = \tau + \{E(\varepsilon_i \mid D_i = 1) - E(\varepsilon_i \mid D_i = 0)\},$$

which is unbiased if and only if disturbances and the treatment are not correlated. Suppose instead that the true model actually equals:

$$Y_i = \alpha + \beta X_i + \tau D_i + \gamma_i,$$

where X_i is a matrix of covariates – perhaps representing aptitude in the class – and β is a vector of population coefficients relating aptitude to final exam performance. Assume the disturbances, γ_i, are distributed Normally with zero mean. Consequently $\varepsilon_i = \beta X_i + \gamma_i$, and the difference estimator yields:

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = \tau + \{E([\beta X_i + \gamma_i] \mid D_i = 1) - E([\beta X_i + \gamma_i] \mid D_i = 0)\}.$$

By implication, if D_i and X_i are correlated, then the difference estimator will suffer from omitted variable bias.[5] The prima facie causal effect will over-estimate or under-estimate the true treatment effect depending on whether high-achievers or low-achievers, respectively, are more likely to attend an optional review session.

Two approaches to utilizing covariates in order to achieve *ex post* control are matched-sampling and regression.[6] Matched-sampling models are the counterfactual model's technical counterpart, aiming to mimic controlled experiments by enabling "like with like" comparisons between treated and untreated subjects. For instance, each treated subject can be paired with a comparison group of the most similar untreated subjects using *propensity scores*, which are the estimated probability of receiving a treatment; conditional on the covariates; see Rosenbaum and Rubin (1983). The treatment effect is computed as a weighted average of intra-pair differences. The regression model for exerting *ex post* control is the "difference estimator with additional covariates" (Stock and Watson 2007):

$$E(Y_i \mid X_i, D_i = 1) - E(Y_i \mid X_i, D_i = 0) = \tau + \{E(\gamma_i \mid D_i = 1) - E(\gamma_i \mid D_i = 0)\},$$

where conditional expectations for treatment group and control group, respectively, could be written:

$$E(Y_i \mid D_i = 1) = \alpha + \beta X_i + \tau + E(\gamma_i \mid D_i = 1)$$

$$E(Y_i \mid D_i = 0) = \alpha + \beta X_i + E(\gamma_i \mid D_i = 0)$$

Angrist and Pischke (2009: 80-91) argue that because matching and regression approaches differ only in how they construct weighted averages of the same comparisons between treated and untreated subjects, they will yield similar estimates of treatment effects. Whether these estimates are accurate depends on the conditional independence assumption: conditional on subjects' measured characteristics, treatment selection is independent of potential outcomes. If this condition is met, then the treatment selection bias is "overt," and can be controlled. Rubin (1978) refers to this situation as "ignorable" treatment selection, which Barnow, Cain and Goldberger (1981) refer to as "selection on observables."

3. Covert Bias and Endogeneity

If treatment selection is a function of subjects' unmeasured characteristics – a situation that Barnow et al. (1981) refer to as "selection on unobservables" – then the conditional independence assumption fails and the treatment selection is "unignorable" (Rubin 1978). Figure 1 provides two simple illustrations that feature a treatment-selection stage followed by an outcome-determination stage. In Figure 1A, bias is overt, for the variables that systematically determine treatment selection are observed: X denotes a set of observed variables that affect treatment selection and the outcome, and Z denotes a set of observed variables that only affect treatment selection. Idiosyncratic variation in treatment assignment is attributed to one disturbance term, e_1, and idiosyncratic variation in outcomes is attributed to another disturbance term, e_2. In Figure 1B, bias is covert, for a set of unobserved variables, denoted W, systematically affect outcome and treatment selection.[7]

Figure 1 about here.

Consider again the optional review session analogy. Elements of X are measureable variables that affect students' willingness to attend and their likely achievement on the final exam, such as taking the course in their major or performance on a practice test. Elements of W can be conceived of as variables that affect students' likely performance and their willingness to attend a review session but are difficult to measure. If X represents the observed aptitude for a course, such as homework scores, then W represents unobserved aptitude, such as test-taking skills. (Students who have a history of testing well may be less

likely to attend review sessions due to greater confidence in their own abilities.) Z represents factors that affect *only* the convenience of attending a review session, such as how close to campus a student resides, or whether a student is employed in a part-time, off-campus job.

We simulated 1000 observations based on Figure 1B using a latent variable framework. W, X and Z are normally distributed random variables, independently drawn from populations with means 0 and standard deviations 1. Disturbances, e_1 and e_2, are normally distributed random variables, independently drawn from populations with means 0 and standard deviations 1. We generate D* and Y as follows:

$$D_i^* = (-1)\cdot W_i + (1)\cdot X_i + (1)\cdot Z_i + e_{1i}$$

$$D_i = 1 \text{ if } D_i^* > 0$$
$$D_i = 0 \text{ if } D_i^* \leq 0$$

$$Y_i = (-1)\cdot W_i + (1)\cdot X_i + (1)\cdot D_i + e_2$$

D* is a latent variable representing students' continuously-varying willingness to attend the review session, and D is the observed decision of whether to attend, based on surpassing a threshold (set equal to zero). The true effect of the binary treatment (D) is $\tau = 1$. The observed variables (X) positively affect selection and the outcome, and the unobserved variables (W) negatively affect selection and the outcome.

We employ four separate estimators of treatment effect. First, we estimate the naïve observed average treatment effect (OATE) by subtracting the mean Y for absentees from the mean Y for attendees. Second, we use the "difference estimator with additional covariates" to estimate the treatment effect as the coefficient in an ordinary least squares regression of Y on D and controls. Third, we estimate the average treatment effect on the treated (ATET) using nearest-neighbor propensity-score matching. We estimate a probit regression of D on controls, then use this selection equation's predicted values to pair each treated subject with the most similar untreated subject; differences in the paired outcomes are computed and aggregated by weighted averaging. Fourth, we use a simplistic instrumental variable least squares regression. We estimate a probit regression of D on Z and controls, and then use the predicted probability of D as an instrumental variable in a regression of Y on the controls, excluding Z.

We estimated the treatment effect four ways under two specifications: W included (in both first- and second-stages) and W omitted (from both stages). If residuals are generated under the assumption that W is unobserved (i.e., $u_{1i} \equiv D_i^* - X_i - Z_i$ and $u_{2i} \equiv Y_i - D_i^* - X_i$), they correlate at 0.386; consequently omission of W will lead to an *overestimate* of the true treatment effect. The 95% confidence intervals for the four methods, for the two specifications, are as follows:

	W : Included	W : Omitted
Naïve Difference Estimator	[2.511, 2.902]	
Ordinary Least Squares	[0.833, 1.138]	[1.771, 2.143]
Propensity-Score Matching	[0.475, 1.437]	[1.860, 2.445]
Instrumental Variable Least Squares	[0.553, 1.066]	[0.269, 1.138]

If W is included in the specification, then the bottom three models' confidence intervals include the true value of the treatment effect. Ordinary least squares generates the narrowest confidence interval among the three. If we omit W from the specification, then only the instrumental variable model's confidence interval contains the treatment' true value ($\tau = 1$); the other three methods overestimate the treatment effects by large amounts. (Appendix Table 1 presents the full estimates from the first- and second-stage models, plus a Heckman-type treatment regression not reported in the text, for both specifications.)

This simple simulation provides three insights. First, both propensity score matching and regression suffer from covert bias when treatment selection depends on unobservable and/or unmeasured variables. Because the unobserved characteristic has parallel effect on likely selection and the likely outcome, the prima facie causal effect overestimates the treatment effect. Second, an instrumental variables model can mitigate the bias in the selection-on-unobservables context, as long as an exclusion restriction (Z) can be found. Instrumental variables modeling is just one possible cure to endogeneity problems with a binary treatment; we present more options later in the paper. Third, three estimators generate similar estimates of the treatment effect when D is exogenous (i.e., W included), but generate differing estimates when D is endogenous (i.e., W omitted). This third insight plays a crucial role in how one approaches the task of diagnosing endogeneity, as we discuss in the next section.

4. Diagnosing Endogeneity Problems

How can the practitioner be confident that the analysis is not tainted by endogeneity problems? This difficulty is particularly salient when one must address a concern raised by skeptical or suspicious audiences or readers (e.g., journal peer reviewers). Because OLS generates residuals that automatically are random disturbances, they will be uncorrelated with regressors (any correlation will be subsumed by the estimated coefficients, producing bias) and with the residuals from another equation. Hence, one cannot simply examine the residuals to diagnose endogeneity.

The first step in addressing the problem should always be theory, asking: How likely is it that the treatment variable is fixed in repeated samples? Macro-economists treat weather as an exogenous factor, and political scientists treated gender as an exogenous factor (until modern science intervened, that is). If the treatment variable is not fixed in repeated samples, the second step in addressing the problem is to ask: How well do the covariates explain treatment selection? Answering this question requires the practitioner to follow Achen's advice to model "both the behavioral outcome of the experiment *and* the assignment to treatment groups" (1986, 37). The difficulty is that there are no clear guidelines on interpreting treatment assignment results. Consider a hypothetical: some scholars have claimed that allowing young children to watch television lessens their attention span. If a first-stage model has extremely strong fit, because measureable demographic factors predispose children both to the treatment (television viewing) and the outcome (attention span), then the causal claim most likely would be undermined if the selection effect was controlled. But, a first-stage model might have extremely poor fit for two, opposite reasons: because the primary determinants of assignment are unobserved factors, such as parental attitudes or care-giving tendencies, *or* because television viewing is an unpredictable phenomenon, and can be treated as if it is randomly assigned. Simply examining propensities to view television does not offer a true escape from the predicament.

Addressing the problem requires using the results of the selection model as a mechanism for forming comparisons between estimators. For instance, matched-sampling estimators prove their worth by showing covariate imbalance before and after matching. Analogously, instrumental variables

estimators, which assume a treatment is endogenous and attempt to confront the endogeneity directly, can prove their worth through a comparison to a baseline estimator that assumes a treatment is exogenous. The simplest version of this comparison has three elements. The researcher estimates an OLS model, which is consistent if the null hypothesis of endogeneity is true, and further, is guaranteed to be efficient whether the null hypothesis of exogenity is true or false (Maddala 1983). The researcher also estimates an instrumental variable model, which is consistent if the null hypothesis of exogeneity is false. Third, one assesses whether the vector of estimated coefficients differs by more than just sampling error; the test statistic is distributed χ^2 with one degree of freedom for each endogenous regressor (see Hausman 1978).[8]

In short, diagnosing endogeneity with binary treatment effects requires both thoughtfulness about endogeneity's potential sources and willingness to estimate multiple models, compare results, and discard findings that are tainted by endogeneity bias. In the next section we turn to a more detailed presentation of a broader set of options available to the researcher.

5. A Menu of Solutions

In an ideal world, regressors and the disturbance term are uncorrelated, and OLS or matching produces unbiased estimates of treatment effects. When this assumption fails, the practitioner needs to be aware of the other options. We offer five solutions, three of which utilize instrumental variables, and the remaining two utilize "control function" approach due to Heckman (see Barnow et al. 1981; Heckman and Robb 1985).

Instrumental variables (IV) methods for binary endogenous regressors are adapted from the two-stage least squares (2SLS) model – the solution offered by economists since the Cowles Commission. The basic idea is to find a source of autonomous variation (see Aldrich 1989); incorporating this variation can purge regressors of their correlation with the disturbance term. If the endogenous variable is continuous, then the first stage would be a regression of the endogenous variable on a set of exogenous variables, including "purging variables" that (are conjectured to) affect the treatment but not the outcome. The second stage would be a regression of the outcome variable on the exogenous variables, minus the

purging variables, plus the predicted values from the first-stage regression as substitute for the original values of the endogenous variable. Assuming that the instruments are valid, this process produces consistent estimates. But if a possibly endogenous treatment variable is binary, this complicates the process of creating an instrument, for the same reasons that a binary dependent variable complicates regression analysis (see Aldrich and Nelson 1984).

The first IV approach to consider, one that requires the smallest modification to 2SLS, uses a linear probability model (LPM) to estimate the first stage equation, using Goldberger's (1964) corrections to standard errors. The predicted probabilities calculated from the first stage are then used in the second stage, after resetting any out-of-bounds estimates. Achen (1986: 44-45) refers to this process as "generalized two-stage least squares" (G2SLS). Achen uses this method to show that pretrial detention actually has no effect on increasing the probability of a conviction, once the analyst properly controls for the nonrandom assignment of arrested individuals to detention or release (see Schneider et al. (1997) for another application). Proponents of this model contend that because the first-stage equation is a linear model, coefficients in the selection equation can be interpreted straightforwardly as changing the probability of treatment.

A second IV model uses nonlinear regression for the first stage equation. Achen (1986: 46-48) describes a handful of nonlinear methods, such as including squared and cubed terms, but the model that has most in common with propensity score matching and other methods we discuss below uses probit to estimate the treatment selection coefficients. This is one version of what Alvarez and Glasgow (1999) label two-stage probit least squares (2SPLS).[9] The predicted values from the first-stage probit model can be substituted for the actual values of the treatment variable in the second-stage least squares model. Angrist and Pischke (2009: 190-1) warn against using this method, for if the first-stage is not probit (i.e., if disturbances in the treatment-selection stage are not distributed Normal), then there is no guarantee that first-stage residuals will be uncorrelated with fitted values of the endogenous treatment and covariates.

A third IV model, proposed by Wooldridge (2002: 623-625) and endorsed by Angrist and Pischke (2009: 191), inserts an intermediate stage, and hence we refer to the procedure as the three-step approach.

Step 1 is a probit regression of the endogenous dummy variable on the exogenous variables and the exclusion restrictions. Step 2 is a least squares regression of the endogenous treatment variable on the exogenous variables and the predicted probabilities from Step 1. Step 3 is a least squares regression of the outcome variable on the exogenous variables and the predicted values from Step 2. The procedure begins no differently from the probit model of selection, and hence exclusion restrictions must be found. The second step uses first-step predicted probabilities as its exclusion restrictions; the intermediate step allows the researcher to employ a non-linear probability for the assignment of the treatment but does not impose a specific distributional assumption for the probability model.

A salient concern with all IV methods is that the system of equations must be identified in order to obtain estimates of the parameters. Since many explanatory variables appear in both the selection and outcome equations, there can be a high degree of multicollinearity between the instrumental variable and the covariates. These IV procedures may be technically identified even if both equations employ the same variables, since the predicted values from the selection equation enter the outcome equation in a nonlinear form.[10] Exclusion restrictions, or purging variables, are notoriously difficult to find, since they must be "exogenous," meaning that they do not affect the outcome, and "relevant," meaning that they do affect treatment selection. If the exclusion restrictions weakly influence selection, IV methods can be badly inconsistent even in large samples (Bound, Jaeger, and Baker 1995). If the exclusion restrictions do indeed affect the outcome, i.e., are quasi-instruments, then instrumental variables methods may be plagued by excessive mean-squared errors (Bartels 1991). In either situation, i.e., weak instruments or quasi-instruments, the practitioner might do better to ignore the endogeneity problem and employ OLS. We discuss tools for diagnosing instrument validity in the conclusion.

Two alternatives that do not employ instrumental variables are "control function" (CF) models based on the work of Heckman (1978, 1979, 1990, 2000).[11] The Heckman models can be estimated via a two-step approach or by Full-Information Maximum Likelihood (FIML) methods; the latter has efficiency advantages in the presence of selectivity bias (Hartman 1991), but is theoretically less robust with respect to specification errors because it assumes bivariate normality of the errors (Alvarez and Glasgow 1999).

Heckman's idea is to estimate a model of the treatment assignment, and then use the predicted probability of treatment as a separate covariate, called a control function or a hazard function. Let us illustrate this approach using the two-stage method, in which the first stage is a probit model of treatment selection, using the familiar latent variable formulation:

$$D_i^* = Z_i\alpha + u_i$$

$$D_i = 1 \text{ if } D_i^* > 0$$
$$D_i = 0 \text{ if } D_i^* \leq 0,$$

where Z is a vector of exogenous characteristics that influence whether the observational unit receives the treatment, α is a vector of coefficients to be estimated, and D^* is a continuous, latent variable for the binary, observed treatment variable D. Further, let Y denote the outcome of interest, let X denote a vector of variables that explain the outcome according to the coefficients β:

$$Y_i = X_i\beta + \tau D_i + \gamma_i.$$

The error terms from the selection equation (u_i) and outcome equation (γ_i) can be correlated. If we assume errors are generated from a bivariate Normal distribution, then we can derive the following regression equation for treated subjects (see Greene 2003, 787-788):

$$\begin{aligned} E[Y_i \mid D_i=1,\ X_i] &= X_i\beta + \tau + E[\gamma_i \mid D_i=1] \\ &= X_i\beta + \tau + \rho\sigma_e \left[\frac{\varphi(Z_i\alpha)}{\Phi(Z_i\alpha)}\right], \end{aligned}$$

where φ denotes the standard normal density function, Φ denotes the standard cumulative normal distribution function, ρ is the correlation between the disturbances in the selection and outcome equations, and σ_e is the standard deviation of the errors in the outcome equation. We can also derive the regression equation for untreated subjects:

$$\begin{aligned} E[Y_i \mid D_i=0, X_i] &= X_i'\beta + E[\gamma_i \mid D_i=0] \\ &= X_i'\beta + \rho\sigma_e \left[\frac{-\varphi(Z_i\alpha)}{1-\Phi(Z_i\alpha)}\right]. \end{aligned}$$

The treatment effect is defined as the following difference:

$$E[Y_i \mid D_i=1, X_i] - E[Y_i \mid D_i=0, X_i] = \tau + \rho\sigma_e \left[\frac{\varphi_i}{\Phi_i(1-\Phi_i)}\right],$$

where $\varphi_i=\varphi(Z_i'\alpha)$ and $\Phi_i=\Phi(Z_i'\alpha)$. The consequence of using OLS to estimate the treatment effect is indicated by the value of the second term on the right-hand-side of this equation. The OLS estimate of τ will be unbiased only if $\rho = 0$. Because σ_e and $\frac{\varphi_i}{\Phi_i(1-\Phi_i)}$ are always positive, the sign of ρ indicates the direction of the bias. If $\rho > 0$ then ordinary least squares will overestimate the treatment effect; if $\rho < 0$ then regression will underestimate the treatment effect.

Heckman's solution is to incorporate information about treatment selection into the outcome equation by adding a regressor, a selectivity-correction variable, $\hat{\lambda}_i$, defined as follows:

$$\hat{\lambda}_i = \frac{\varphi(Z_i\hat{\alpha})}{\Phi(Z_i\hat{\alpha})} \qquad \text{if } D_i=1,$$

$$\hat{\lambda}_i = \frac{-\varphi(Z_i\hat{\alpha})}{1-\Phi(Z_i\hat{\alpha})} \qquad \text{if } D_i=0.$$

If we assume that the joint distribution of the error terms in the selection and outcome equation is bivariate normal, estimates of α can be obtained from a probit model estimated via a maximum-likelihood procedure,[12] and estimated values of λ can be calculated and included in the second-stage OLS regression.

The Heckman FIML model provides a direct test for endogeneity through a test of the significance of □. The Heckman two-step procedure provides an indirect test for endogeneity, estimating □'s statistical significance through the estimated coefficient on the selectivity-correction variable ($\hat{\lambda}_i$). In fact, all five techniques that we have discussed provide a method for testing for endogeneity of the treatment. For the three IV models, one can use the (Durbin-Wu-) Hausman test for endogeneity described in Section 4. If a researcher chooses to confront the endogenous treatment selection problem, then immediately after estimating any of these models, the next step should be assessing evidence of endogeneity, since OLS is more efficient if the treatment is (conditionally) exogenous.

With numerous options available for confronting endogeneity problems, it worth assessing how these procedures perform under different conditions. In the next section we report on simulations that

investigate the estimators' performance under varying levels of endogeneity, varying levels of instrument strength, and varying distributions of the errors in the first-stage treatment equations.

6. Monte Carlo Simulations

In order to examine the properties of these estimation procedures, we performed Monte Carlo simulations. The sample size for each simulation is 1000 observations. We created 28 separate conditions from varying the distribution of the disturbance term in the treatment-selection equation, the degree of endogeneity (i.e., correlation between the disturbances in the two equations), and the strength of the instruments. For each of the 28 conditions, we repeated the simulation 1000 times. After every simulation we computed an estimate of the treatment effect using ordinary least squares and each of the five estimation procedures discussed in Section 4.

The underlying model is as follows:

$$d^* = \alpha z + 0.5x + e_1$$

$$d = 1 \text{ if } d^* > 0$$
$$d = 0 \text{ if } d^* \leq 0$$

$$y = 1d + 1x + e_2$$

The treatment effect is $\tau = 1$, as seen in the outcome equation. The parameter α in the treatment-selection equation determines the relevance of the instrument, i.e. the strength of the relationship between the instrument and the endogenous variable. Each equation has an associated disturbance: e_1 is the error term for the treatment equation and e_2 is error term for the outcome equation. The error terms are generated in the following manner:

$$e_1 = F(0,1)$$

$$e_2 = N(0,1) + \delta e_1$$

where $F(\bullet)$ is a generic probability distribution and $N(\bullet)$ is the Normal distribution. To compare the different procedures and to examine their robustness, we utilize three different functional forms for $F(\bullet)$. First, we generated treatment-selection disturbances using a Normal distribution. The Heckman FIML model, the Heckman Two-Step model, and the 2SPLS model all employ a probit model in the first stage,

therefore the normality assumption favors these models. However, incorrect distributional assumptions might have disastrous effects, particularly for the FIML model. Therefore, we generated treatment-equation errors using the logistic distribution, which is symmetrical and bell-shaped like the Normal distribution, but with fatter tails. We chose the logistic distribution since it is similar but not identical to the Normal distribution, and thus represents a small departure from normality. Following Vella and Verbeek (1999), who investigate models with endogenous treatments effects, we also generated first-stage errors using a uniform distribution, which represents a large departure from Normality.

Regarding the degree of endogeneity, we generate e_2 as a weighted combination of e_1 and a standard Normal random variable. We chose seven distinct values for the weight, δ, to produce seven different levels of endogeneity, producing correlations between the two equations' errors (denoted $\rho_{e1,e2}$) that range from 0.0 to 0.95. The error terms were standardized to have a mean of 0 and a variance of 1 before parameter estimations were carried out.

As we discussed above, invalid instruments can undermine the value of performing IV analysis, particularly because instruments that are truly autonomous may be only weakly related to the treatment selection. To examine the properties of the estimators under varying levels of instrument relevance, we performed simulations using both strong- and weak-instrument conditions.[13]

In rare instances, the FIML model failed to converge in the Monte Carlo simulations using the default optimization algorithms. Generally speaking, this is more likely to occur when the degree of endogeneity is high (i.e. the error correlation is 0.9 or greater), when the instrument is weak, and when the sample size is small.[14] We minimized this problem's occurrence by holding the sample size at 1000, but closed-form solutions for the first and second derivatives of the log-likelihood function do not exist for the Heckman FIML procedure. In such a situation, when convergence is not assured, the researcher has a choice of updating algorithms that are asymptotically equivalent. Following Eliason (1993), we set the optimization sequence to 50 iterations under the Broyden, Fletcher, Goldfarb, and Shannon (BFGS) algorithm to get close to the solution, then switch to 5 iterations under the Berndt, Hall, Hall, and

Hausman (BHHH) algorithm, which is more efficient. This sequence was repeated until convergence was achieved, which happened in less than 500 iterations in every one of the simulations reported below.

To compare the properties and performance of each procedure, and to evaluate the consequence of changing levels of endogeneity and instrument relevance, we report and analyze three criteria: average bias, mean-squared error, and coverage.

The bias of an estimate is the difference between the true value of the treatment effect ($\tau = 1$) and the estimated treatment effect. In Table 1 we report the average bias in treatment effect estimates, across 1000 simulations for each condition. As a baseline, consider the average bias of ordinary least squares (shaded rows), noting that the distribution of the treatment-selection stage error term is irrelevant for the OLS procedure, since treatment-selection is not part of the estimation procedure. OLS is, on average, barely biased when there is no endogeneity problem ($\rho_{e1,e2} = 0$). As the level of endogeneity increases ($\rho_{e1,e2}$ increases moving across the columns, from left to right), OLS becomes more biased. To interpret the bias figures, observe that when the disturbances are correlated at 0.1, the average bias is between -0.147 and -0.167, indicating that the average estimated treatment effect is 15% to 17% larger than the true treatment effect. When the correlation reaches 0.7, the average bias is over 100% of the true value for any of the three functional forms (Normal, logistic, or uniform).

INSERT TABLE 1

How do the other estimators perform? As long as the errors are drawn from a Normal distribution and the instrument is strong, the average bias never exceeds 2% for the Heckman Two-Step model, 2SPLS, Three-Step, or Heckman-FIML. Average bias of the LPM model is higher, but still never exceeds 10% of the true treatment value. When we compare the results for strong and weak instruments – maintaining the assumption of Normal errors – we observe that average bias is consistently greater for the three IV models than for the two CF models. The Heckman Two-step method performs well across the board, consistently yielding less average bias than the IV models. The Heckman-FIML model performs poorly for medium levels of endogeneity, but performs excellently for both high and low levels of

endogeneity. The 2SPLS models performing slightly better than the Three-Step model in the weak instrument condition, and the LPM performs consistently worst, as usual.

The average biases of the estimates drift further apart once the true data generating process deviates from Normality. When errors are generated from a Logistic distribution, the average bias for the Three-Step method or the 2SPLS model is always lowest. In fact, a comparison of the results for the Normal and Logistic functional forms reveals little change in the Three-Step method's average bias under different, indicating its superior robustness. The two CF approaches yield more biased estimates than the Three-Step model and 2SPLS, especially at higher levels of endogeneity (higher values of $\rho_{e1,e2}$). The average bias in the Heckman Two-step estimates is small, about 3 percent of the true value of the treatment effect if the correlation between the errors is large. The average bias of the Heckman-FIML estimates is roughly three times as large as the Two-Step model's bias when errors are generated with a logistic functional form. The dangers of imposing the wrong distributional assumption become most evident when we examine the bias with uniform distributed errors. Although the average bias worsens for all methods, the Heckman Two-Step model yields less biased estimates than the Heckman-FIML model.

To summarize the discussion of bias, we find first that the level of endogeneity has barely an effect on the two non-linear IV models (2SPLS and Three-Step) and on the two CF models when the errors are drawn from a Normal functional form and the instrument is strong. When the instrument is weak, the Heckman Two-Step model is less biased than the two non-linear IV models, and its superiority becomes more apparent as the level of endogeneity grows. However, the distribution used to generate the errors interacts with the level of endogeneity, generating substantially greater bias when the errors are generated from distributions with thicker tails (Logistic or uniform). The Heckman-FIML model's bias increases most dramatically with increasing endogeneity if the Normality assumption is violated.

A second criterion for judging among estimators is the mean-squared error (MSE), which is the sum of the squared differences between the estimated and true treatment effect. MSE also equals the sum of variance and squared bias, so the least biased estimators are expected to perform better on this criterion as well. We present the average mean-squared error for the different conditions in Table 2. As the

baseline for comparison, the OLS simulation results show that as the correlation between the error terms increases, the MSE far exceeds the other approaches. If the expected correlation is very low (i.e., less than or equal to 0.1) and/or the instruments might lack validity, then OLS may be advisable, as Bartels (1991) has suggested. However, if we are suspicious of endogeneity but confident in the instruments, then almost any method for controlling endogeneity is an improvement over OLS. The MSE results indicate that the Heckman FIML model yields a lower MSE than any alternative approach if the first stage errors are drawn from a Normal or Logistic distribution. Under the Normal functional form, this was to be expected, given that maximum likelihood estimation is efficient and consistent, and it is reassuring that the Heckman-FIML model's MSE is still lowest when the true data-generating process follows the Logistic model. If the true data-generating process is Uniform, the MSE of the Heckman Two-Step approach is lowest if the correlation is 0.9 or greater. The Heckman Two-Step method and the other three nonlinear models yield similar MSE values to each other regardless of the correlation between the errors. Based on a MSE criterion, the FIML approach appears superior (except when the level of endogeneity is low, when OLS may be the preferred approach); the superiority is even greater for the weak instrument and Logistic-distributed errors conditions than with Normal errors and strong instruments.[15]

INSERT TABLE 2

A third criterion for choosing among estimators is the coverage rate, defined as the percent of simulations for which the true value located in the estimated 95% confidence interval. We present the coverage rates for each of the six estimation procedures in Table 3. Naturally, we expect that the true value of the treatment effect should fall in the estimated confidence interval 95% of the time. For OLS, we see that the coverage rate is close to 95% when there is no endogeneity problem, but quickly falls as endogeneity rises. When the correlation between the disturbance terms is 0.3 or greater, the coverage rate for OLS is 0% for the 15,000 simulated samples. Under the assumption of normality, we find that the coverage rate is close to 95% for the other methods. The coverage rate for the LPM approach is consistently greater than 95% regardless of the level of endogeneity or distributional assumption. How can it be that the model with the worst bias has the best coverage? The simple answer is that the

confidence interval generated by the LPM is too large. The Heckman FIML approach looks slightly less appealing under Normal errors, and its coverage rate decreases as the level of endogeneity increases for the logistic or uniform distributions. The other methods are more promising, in particular the Heckman Two-Step model and the Three-step model. The 2SPLS model yields coverage rates that are consistently too high, while coverage rates for the Heckman Two-Step model and the Three-step model are consistently close to 95%, for all levels of endogeneity and distributional assumptions.

INSERT TABLE 3

In the final section of the paper we offer guidance on which method – or *methods* – would be best depending on the circumstances the researcher confronts. Based on the evidence on bias, mean-squared error, and coverage, the simulations guide us to endorse the Heckman Two-Step or the Three-Step IV model, assuming the researcher faces an endogeneity problem. These procedures have consistently low MSE and appropriate coverage rates. The Heckman Two-Step procedure generates slightly more biased estimates if errors at the treatment-selection phase do not follow a Normal distribution, but is more robust to weak instruments than the IV approaches. The Three-Step IV approach has several notable advantages, particularly robustness to misspecification, and the ease with which it can be extended to multiple endogenous variables, some of which may be continuous and some of which may be dichotomous. The Heckman FIML procedure is least biased and has lowest MSE when the true data generating process is Normal, but we hesitate to recommend it because of its sensitivity to misspecification. When the errors are generated from a non-Normal process, the Heckman-FIML approach yields biased estimates and a low coverage rate, especially as endogeneity increases.

7. Application to Presidential Leadership in Congress

To demonstrate the consequences of using these different procedures, we replicate the analysis of Canes-Wrone (2001) who examines whether presidents' "going public" impacts success in Congress.[16] Canes-Wrone measures success by calculating the proportional difference between a president's requested appropriation and Congress's actual appropriation for various government agencies. Dividing the raw dollar amounts by the agency's budget from the previous fiscal year ensures that success is

comparable across agencies of various sizes. Canes-Wrone theorizes that the binary treatment, a public appeal by the president, is endogenous to success because presidents are strategic in selecting which agencies to emphasize. Presidents will not wish to be perceived as foolishly wasting their political capital on an unpopular proposal, or on a proposal that is already likely to pass. If a president is more apt to make a public appeal when it is likely to bend Congress to the president's will, then factors affecting success will also affect appeals. A practical reason for replicating this analysis is that Canes-Wrone (2001) provides an instrument that passes both tests of validity: the size of the agency, measured by the previous year's budget. First, with regard to instrument strength, presidents are more likely to allocate their scare time to larger and hence more salient agencies, all else being equal. In the results reported in Table A2, the coefficient for agency size is statistically significant ($p < 0.05$) in every specification except in the LPM (where $p = 0.07$).[17] Second, with regard to instrument exogeneity, notice that the dependent variable is the *proportional* change in the request, which is, by construction, unaffected by agency size. An added practical reason to replicate Canes-Wrone's (2001) research is that the analysis utilizes more than 1000 observations, providing a large enough sample to estimate the various procedures discussed in this paper.[18]

It is important to note that Canes-Wrone (2001) estimates a simultaneous equation model with two jointly endogenous variables: public appeals and presidential budgetary success.[19] Simultaneous estimation of these two equations may be more efficient, but two equations can be estimated separately without compromising consistency. Given our interest in models for estimating a dichotomous treatment variable's effect on a continuous response variable, we restrict our attention to the budgetary success equation. The dependent variable is the *negative* absolute value of the difference between the president's requested appropriation and Congress's final appropriation; these figures are normalized by dividing by the previous year's appropriation. The treatment of interest is a public appeal, measured using a dummy variable coded equal to one if the president made a statement about the agency's budget in a nationally televised address, and coded equal to zero otherwise. Numerous control variables are specified in the

model; we present the full models in the appendix and urge the reader to consult Canes-Wrone (2001) regarding justification and coding conventions for the control variables.

Canes-Wrone predicts and finds a positive treatment effect: a presidential appeal leads to greater presidential success (Table 2; 2001: 325). We wonder, does the estimated treatment effect vary across estimation procedures? Table 4 presents the estimated treatment effects, along with estimated standard errors and *p*-values, for least squares regression and matched-sampling (which assume conditional exogeneity) and the five alternative methods (which assume endogeneity). The point estimate and 95% confidence interval estimates are also displayed in Figure 2. Table 4 also reports the appropriate *p*-value from a test of whether the treatment variable is actually endogenous. For the Heckman-FIML model, the proper endogeneity test is a likelihood ratio test of the two equations' independence, with $\rho = 0$ being the null hypothesis. For the Heckman Two-Step model, the proper endogeneity test is a test of the hypothesis that the control function coefficient (λ) equals zero. For the IV models, the proper test of endogeneity is a Hausman test. In all cases, the null hypothesis is that the treatment variable is exogenous.

The first two rows of Table 4 present treatment effect estimates using Ordinary Least Squares and Nearest-Neighbor matching (nearest *k*-neighbor matching with bootstrapped standard errors, where k=3). The OLS estimated coefficient is positive, consistent with Canes-Wrone's theory, but the magnitude of the coefficient is small ($\beta = 0.057$) relative to the standard deviation of the dependent variable (0.190). Matching produces a slightly larger estimate of the treatment effect ($\beta = 0.066$). Both estimated treatment effects are statistically significant at conventional levels.

The other five methods produce larger estimates of the treatment effect than the OLS or matching methods. The Heckman-FIML procedure produces an estimate of the treatment effect that is only slightly larger ($\beta = 0.082$) and this coefficient is significantly different from zero, however the likelihood ratio test of endogeneity fails to reach the conventional level of statistical significance ($p = 0.47$). The Heckman-FIML results thus suggest that controlling for endogeneity is not necessary and the OLS estimates are valid. As we have discussed, however, FIML models lack robustness if the model is mis-specified. To use

the Heckman model appropriately, it must be the case that arrows between the outcome variable, presidential success, and the treatment variable, presidential appeals, do not run in both directions. Canes-Wrone (2001) theorizes that causality is reciprocal, and if she is correct, then the FIML estimates must be invalid. This critique *only* applies to the FIML estimates, because the models utilizing two or more steps (i.e., Heckman two-step, 2SPLS, Wooldridge's three-step method, and linear probability models) do not require that causality is unidirectional.

The Heckman Two-step estimate of the treatment effect is quite large ($\beta = 0.234$) relative to the OLS and Heckman-FIML estimates, and it is statistically significant ($p = 0.03$). The test for endogeneity falls in a grey area regarding statistical significance ($p = 0.09$). The 2SPLS model and the Three-Step model also produce estimates of the treatment effect that are similar in magnitude, more than three times as large as the OLS and matching estimates, but they fall short of conventional levels of statistical significance ($p = 0.11$), and the endogeneity tests for these models are insignificant ($p = 0.21$).

Finally the Linear Probability Model produces an estimate that of the treatment effect that is far larger than any of the other models ($\beta = 1.093$) and that approaches conventional levels of statistical significance ($p = 0.08$). Most likely, this large coefficient results from fitting a linear model to a rare event process. Out of 1,124 observations, the president made a public appeal only 88 times, which amounts to a probability of a public appeal below 7 percent. In addition to providing the largest estimated treatment effect, the LPM procedure also provides the most statistically significant test statistic suggesting that endogeneity is a problem; the Wu-Hausman test's *p*-value is less than 0.01. But given the availability of other suitable methods, discussed herein, and given the results from the Monte Carlo simulations, the use of a LPM model with dichotomous treatment variable is hard to justify.

Our aim in replicating this analysis was not to question Canes-Wrone's findings, but rather to utilize available data to demonstrate that treatment effects estimates are sensitive to the chosen method. The results are consistent in finding that presidential appeals have a positive effect on presidential success, as hypothesized, and most effects are larger than Canes-Wrone estimated.[20]

8. Discussion

Given the prevalence of observational, non-experimental data in social science, researchers must be constantly wary of endogeneity problems, which can result from omitted variables, measurement error in independent variables, reciprocal relationships, and non-random treatment assignment. While methods for confronting endogeneity have proliferated, an understanding of which methods are appropriate under which conditions remains elusive. In this article, we have surveyed methods for confronting endogeneity bias in a specific yet common case, when the potentially endogenous treatment is a dichotomous variable and the outcome of interest is continuous. Based on the results of Monte Carlo simulations and a test of presidential 'going public,' we offer some practical advice for researchers who suspect that they may have an endogenous binary treatment variable. We suggest a practitioner follow a seven-step process.

1. Pause, and consider sources of a possible endogeneity problem

The first step should be to push back from the keyboard and consider carefully the source or *sources* of suspected endogeneity. (Fortunately, journal peer reviewers and brownbag audiences often are willing, or even *delighted*, to share their suggestions about probable sources of endogeneity.) If the endogeneity problem originates in non-random assignment of a binary treatment, then the practitioner should first consider if assignment is based on observable or unobservable factors. If the latter, then an IV or CF approach might be worth considering – although not necessarily worth adopting unless other conditions are satisfied. The next steps require the researcher to estimate multiple models, and utilize the results of multiple estimations to identify the proper model.

2. Estimate ordinary least squares regression
3. Estimate instrumental variables regression
4. Perform a Hausman test of endogeneity
5. Perform a test of instrument strength

Ordinary least squares estimation of a system of equations is guaranteed to be efficient regardless, and consistent under the null hypothesis of exogeneity (Maddala 1983). An instrumental variable model will be inefficient, but will be consistent under the alternative hypothesis of endogeneity. In our opinion,

the researcher should estimate two models: OLS and Wooldridge's (2002) Three-Step IV model. Comparing the estimated coefficients on the treatment variable from the OLS and IV models is worthwhile – and we show such a comparison in Figure 2 – but it is only a heuristic device. The fourth step should be an examination of the estimated coefficients; if the null hypothesis of exogeneity is correct, then the two vectors of coefficients should differ only by sampling error. The Hausman test, described earlier, provides a straightforward method for testing the null hypothesis of exogeneity; rejection of the null hypothesis indicates that OLS is inconsistent. However, this result would not imply that an OLS estimate of the treatment effect is inferior to an IV estimate, for an IV model is only as good as its exclusion restrictions. Instrument strength can be assessed a variety of ways in a canonical 2SLS model, but options are limited with a binary endogenous variable. Two options for assessing instrument strength are the statistical significance of the purging variables in the first-stage model and the LR test statistic from the joint hypothesis test of excluding the purging variables from the first-stage probit equation. If the null hypothesis of instrument irrelevance is rejected, and if the Hausman test rejected the null hypothesis of exogeneity, then the IV estimates earn more confidence.

6. Estimate control function regression
7. Perform a Hausman test of endogeneity

In our Monte Carlos simulations, the CF procedures were most robust to weak instruments. This alone makes the two-step Heckman procedure worth considering. Therefore, regardless of instrument strength – because instruments that are too strong might be quasi-instruments – the sixth step is to utilize a CF model, and in particular, Heckman's two-step model, to estimate the treatment effect. The seventh step is another test of endogeneity, examining the statistical significance of the selectivity-correction variable and/or the statistical significance of the estimated ρ. In our replication of Canes-Wrone's analysis, the Three-Step IV model and the Heckman Two-Step model produced nearly equal estimates of the treatment effect, and nearly equal p-values in the tests of endogeneity. We would report all three estimates, which agree that presidential leadership in Congress is effective. Reporting at least OLS and one other estimate would indicate to the reader the size and direction of possible endogeneity bias.

As for the opposite question, of what a practitioner *should not* do, notice that we have not endorsed three of the five models. A FIML model, in which the two-equation model is estimated simultaneously, is sensitive to misspecification. Given that social scientists often must use available data, specification issues present a constant risk, including the difficulties of finding good instruments and measuring concepts. It is concerning to note that FIML estimation is the default in STATA for both the Heckman treatment-selection (*treatreg*) and sample-selection models (*heckman* and *heckprob*). Although the Heckman FIML model has an efficiency advantage vis-à-vis the multi-step models, our Monte Carlo simulations indicate that the FIML version is the least biased under a slim set of circumstances. We also have not endorsed use of an LPM model or the 2SPLS model. Among the IV procedures, the main advantages of the LPM were two: researchers' familiarity with linear regression, and the ease of interpreting first-stage coefficients. As political scientists' comfort level with probit and logit models have grown, however, and as statistical packages have evolved, these advantages have been negated. More importantly, the LPM procedure performed poorly relative to the other two-step methods, producing greater average bias and overly optimistic coverage rates.

More generally, we should re-emphasize that these methods rely on the use of instruments that are strong and autonomous. Other scholars have addressed the issue of the reliability of instruments in the two-stage least squares model with two continuous variables (Bartels 1991; Bound et al 1995). The advice offered in that set of circumstances applies here as well: if the instruments are too weakly related to the treatment, or if the instruments are too strongly related to the outcome, then a model that ignores the endogeneity may be preferred on a MSE criterion, since the standard errors are larger in the two-step procedures without any significant reduction in bias. Unfortunately, there is no real substitute for locating and using valid instruments. So we recommend, finally, that any scholar who draws interpretation from an IV model provide a robust theoretical justification for the validity of the instrument, should consider trying different instruments or combinations of instruments (if available), and should utilize multiple methods to validate the estimates of any individual model.

9. Endnotes

[1] Cox (1988) refers to this as the problem of *non-random assignment*, as distinct from *non-random selection* (which we would refer to as "sample selection bias"). In both situations, the core problem is unobserved data. With non-random selection, no measure of the outcome variable is observed for units who are not in the sample. With non-random assignment, an outcome is observed for all units, but not in both the treated and the untreated states. Bias can result if the treated and untreated units, or if the sampled and unsampled units, differ systematically. See Achen (1986), Dubin and Rivers (1989), and Sigelman and Zeng (1999) on sample selection bias.

[2] Although primarily associated with Rubin (1974, 2006), Holland's (1986) and Rosenbaum's (2002) contributions should not be underestimated. Rubin also gives a great deal of credit to Cochrane. Brady (2008) and Cortina (2009) provide recent and easily digestible introductions.

[3] The bracketed term rarely equals zero *a posteriori*. Quoting Morgan and Winship (2007: 40): "For randomized experiments, the treatment variable is forced by design to be independent of the potential outcome variables. However, for any single experiment with a finite set of subjects, the values of d_i will be related to the values of $y_{i1}$and y_{i0} because of chance variability."

[4] See Manski's (1993: 24) definition of a treatment effect: "the change in average outcome if one were to replace a hypothetical situation in which a person with attributes *X* were exogenously assigned to treatment 0 with another hypothetical situation in which a person with attributes *X* were exogenously assigned to treatment 1".

[5] Stock and Watson urge testing for random receipt of treatment by regressing treatment dummy on covariates, and computing the F-statistic to test whether coefficients are zero. If the treatment is randomly received, then including covariates still would reduce inefficiency (see Angrist and Pischke 2009).

[6] Morgan and Winship (2007) present a matching-heavy approach that is critical of regression, and Angrist and Pischke (2009) present a regression-heavy approach that is critical of matched-sampling. We recommend that the interested reader consult both for balance.

[7] If variables in W affect treatment selection but not the outcome, then they would be classified among e_1; if variables in W affect the outcome but not treatment selection, then they would be classified among e_2.

[8] Obviously, we are aware that Stata performs this as a post-estimation command *estat endog*, following *ivregress*, or as *ivendog*, following *ivreg2* (Baum, Schaffer and Stillman 2003).

[9] Alvarez and Glasgow emphasize 2SPLS in a situation when the endogeneous regressor is continuous and the outcome variable is dichotomous. We refer to this situation as a "dosage selection bias" problem rather than "treatment selection bias," although some readers might view the difference as semantic.

[10] Vella (1998, p. 135) endorses the use of exclusion restrictions, because although the predicted values from the probit model are nonlinear, it is linear in some ranges. Based on the work of Leung and Yu (1996), Vella notes that if there is sufficient variation in some the exogenous variables, then exclusion restrictions may be not be necessary for identification. However, without valid exclusion restrictions, these results should be treated cautiously (see Sartori (2003) for the same point in a different context).

[11] Heckman's motivation was to unite "classical Cowles Commission simultaneous equations theory and models of discrete choice originating in mathematical psychology… in order to produce an economically motivated, low dimensional, simultaneous equations model with both discrete and continuous endogenous variables" (2000: 267). Barnow et al. (1981) deserve credit for the application of Heckman's models to treatment selection problems; they also cite Amemiya (1978) and Maddala and Lee (1976) prominently.

[12] Although economists and statisticians have investigated semi-parametric and non-parametric models of self selectivity, which do not rely on particular distributional assumptions, the more familiar parametric solutions to self selectivity do remarkably well; see Angrist (2001); Heckman (1990); Manski (1993).

[13] In the strong-instrument condition, we set $\alpha \approx .4306$, to achieve a correlation of 0.3 between the instrument (z) and the latent variable (d^*). In the weak-instrument condition, we set $\alpha \approx .1266$, to achieve a correlation of 0.1 between z and d^*. Weakening the instrument has direct effects on the properties of the estimators, however we found no interesting interactions between instrument strength and first-stage equation error term functional forms. Hence, we report simulation results for the logistic and uniform

distributions only for the strong instrument condition.

[14] In a set of simulations with a smaller sample size (*N*=200) that are not reported in this paper, there were a few instances (approximately 1 in every 1000) in which the FIML model failed to converge.

[15] These results parallel Hartman's (1991) findings for the Heckman selection model.

[16] The first year of each presidential administration is excluded, since the prior administration's budget requests are used in the first year, given the timing of the budget process.

[17] The weakness of the instrument in the LPM may contribute to the atypical estimate of the treatment effect reported in Table 4 and Figure 2. Future research could examine what is an acceptable level of instrument strength for these models with a non-linear first-stage equation.

[18] We use the larger of Canes-Wrone's samples (N = 1,124), as the methods we describe are justified by their asymptotic properties. A smaller sample (N = 88) uses the agencies for which public opinion data on the policy area is available.

[19] Canes-Wrone's main interest is whether a policy's popularity and the president's likelihood of success affect the decision to "go public." Researchers who wish to re-examine these hypotheses but who do not wish to estimate the system of equations simultaneously should consult Alvarez and Glasgow (2001), who address model estimation for continuous treatments (a.k.a., *dosages*) and binary outcome variables.

[20] Data provided to us by Canes-Wrone vary slightly from those reported in paper. Notably, the range of the dependent variable (success) is smaller than reported in Canes-Wrone (2001).

10. References

Achen, Christopher H. 1986. *The statistical analysis of quasi-experiments*. Berkeley: University of California Press.

Aldrich, John H. 1989. Autonomy. *Oxford Economic Papers* 41: 15-34.

Aldrich, John H., and Forrest D. Nelson. 1984. *Linear probability, logit, and probit models*. Sage university papers series. Quantitative Applications in the Social Sciences. Vol. 07-045. Beverly Hills: Sage Publications.

Alvarez, R. Michael, and Garrett Glasgow. 1999. Two-stage estimation of nonrecursive choice models. *Political Analysis* 8 (2) (December 16): 147-165.

Amemiya, Takeshi. 1978. The estimation of a simultaneous equation generalized probit model. *Econometrica* 46 (5) (Sep.): 1193-1205.

Angrist, Joshua D. 2001. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics* 19 (1): 2-16.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics : An empiricist's companion*. Princeton: Princeton University Press.

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. 1981. *Issues in the analysis of selectivity bias*. Madison: Institute for Research on Poverty, University of Wisconsin--Madison.

Bartels, Larry M. 1991. Instrumental and "quasi-instrumental" variables. *American Journal of Political Science* 35 (3) (Aug.): 777-800.

Baum, Christopher F., Mark E. Schaffer and Steven Stillman. 2003. IVENDOG: Stata module to calculate Durbin-Wu-Hausman endogeneity tests after IVREG.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association* 90 (430) (Jun.): 443-450.

Brady, Henry E. 2008. Causation and Explanation in Political Science, in Janet Box-Steffensmeier, Henry E. Brady and David Collier (eds). *Oxford Handbook of Political Methodology*. Oxford, UK: Oxford University Press.

Canes-Wrone, Brandice. 2001. The president's legislative influence from public appeals. *American Journal of Political Science* 45 (2) (Apr.): 313-329.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cortina, Jeronimo. 2009. The Potential-Outcomes Model of Causation, in Andrew Gelman and Jeronimo Cortina (eds). *A Quantitative Tour of the Social Sciences*.

Cox, Gary W. 1988. Recent Developments in Statistical Inference: Quasi-Experiments and Perquimans County. *Historical Methods* 21 (Summer): 140-142.

Dubin, Jeffrey A., and Douglas Rivers. 1989. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research* 18 (2-3) (November): 360-90.

Eliason, Scott R. 1993. *Maximum likelihood estimation: Logic and practice*. Sage university papers series. Quantitative Applications in the Social Sciences. Newbury Park, Calif.: Sage.

Goldberger, Arthur S. 1964. *Econometric theory.* New York: J. Wiley.

Hanushek, Eric A., and John E. Jackson. 1976. *Statistical methods for social scientists*. Quantitative studies in social relations series. New York: Academic Press.

Hartman, Raymond S. 1991. A monte carlo analysis of alternative estimators in models involving selectivity. *Journal of Business & Economic Statistics* 9 (1) (Jan.): 41-49.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica* 46 (6) (Nov.): 1251-1271.

Heckman, James J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46 (4) (Jul.): 931-959.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica* 47 (1) (Jan.): 153-161.

Heckman, James J. 1990. Varieties of selection bias. *The American Economic Review* 80 (2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association) (May): pp. 313-318.

Heckman, James J. 2000. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics* 115 (1) (Feb.): 45-97.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15 (3) (June 20): 199-236.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81 (396): 945-60.

Jackson, John E. 2008. Endogeneity and Structural Equation Estimation in Political Science, in Janet Box- Steffensmeier, Henry E. Brady and David Collier (eds). *Oxford Handbook of Political Methodology*. Oxford, UK: Oxford University Press.

Kinder, Donald R., and Thomas R. Palfrey. 1993. *Experimental foundations of political science*. Ann Arbor: University of Michigan Press.

Maddala, G. S. 1983. *Limited-dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Maddala, G.S., and L.F. Lee. 1976. Recursive models with qualitative endogenous variables. *Annals of Economic and Social Measurement* 5: 525-545.

Manski, Charles F. 1993. Identification problems in the social sciences. *Sociological Methodology* 23: 1-56.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference : Methods and principles for social research*. New York: Cambridge University Press.

Rosenbaum, Paul R. 2002. Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* 97 (457): 183-92.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41-55.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5): 688-701.

Rubin, Donald B. 1978. *Using multivariate matched sampling and regression adjustment to control bias in observational studies*. Princeton, N.J.: Educational Testing Service.

Rubin, Donald B. 2006. *Matched sampling for causal effects*. Cambridge; New York: Cambridge University Press.

Sartori, Anne E. 2003. An estimator for some Binary-Outcome selection models without exclusion restrictions. *Political Analysis* 11 (2) (May 01): 111-38.

Schneider, Mark, Paul Teske, Melissa Marschall, Michael Mintrom, and Christine Roch. 1997. Institutional arrangements and the creation of social capital: The effects of public school choice. *The American Political Science Review* 91 (1) (Mar.): 82-93.

Sekhon, Jasjeet. 2008. The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods, in Janet Box-Steffensmeier, Henry E. Brady and David Collier (eds). *Oxford Handbook of Political Methodology*. Oxford, UK: Oxford University Press.

Sigelman, Lee, and Langche Zeng. 1999. Analyzing censored and sample-selected data with tobit and heckit models. *Political Analysis* 8 (2) (December 16): 167-182.

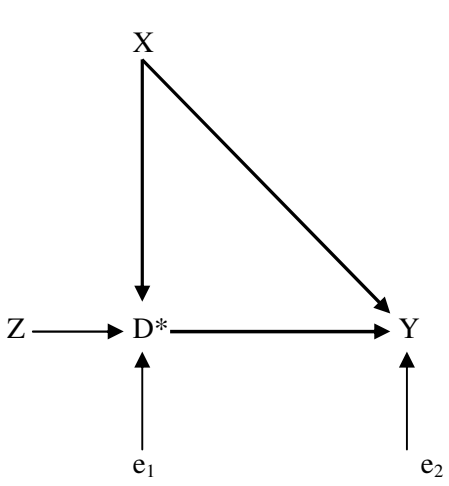
Siu Fai Leung, and S. Yu. 1996. On the choice between sample selection and two-part models. *Journal of Econometrics* 72 (1/2): 197-230.

Stock, James H., and Mark W. Watson. 2007. *Introduction to econometrics*. Boston: Pearson/Addison Wesley.

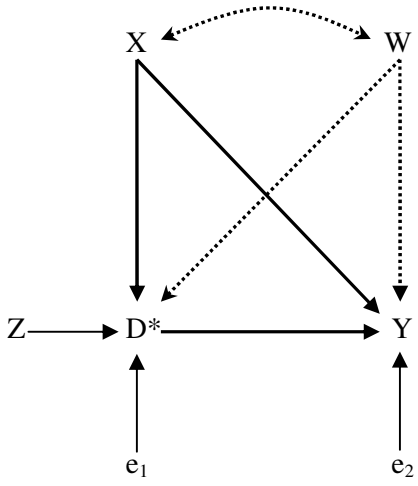
Vella, Francis. 1998. Estimating models with sample selection bias: A survey. *The Journal of Human Resources* 33 (1) (Winter): 127-169.

Vella, Francis, and Marno Verbeek. 1999. Estimating and interpreting models with endogenous treatment effects. *Journal of Business & Economic Statistics* 17 (4) (Oct.): 473-478.

Wooldridge, Jeffrey M.. 2002. *Econometric analysis of cross section and panel data*. Cambridge, Mass.: MIT Press.



(a) Selection on Observables
Ignorable Treatment



(b) Selection on Unobservables
Unignorable Treatment

Figure 1. Causal Diagram

Figure 2: Treatment Effect Estimates of Going Public

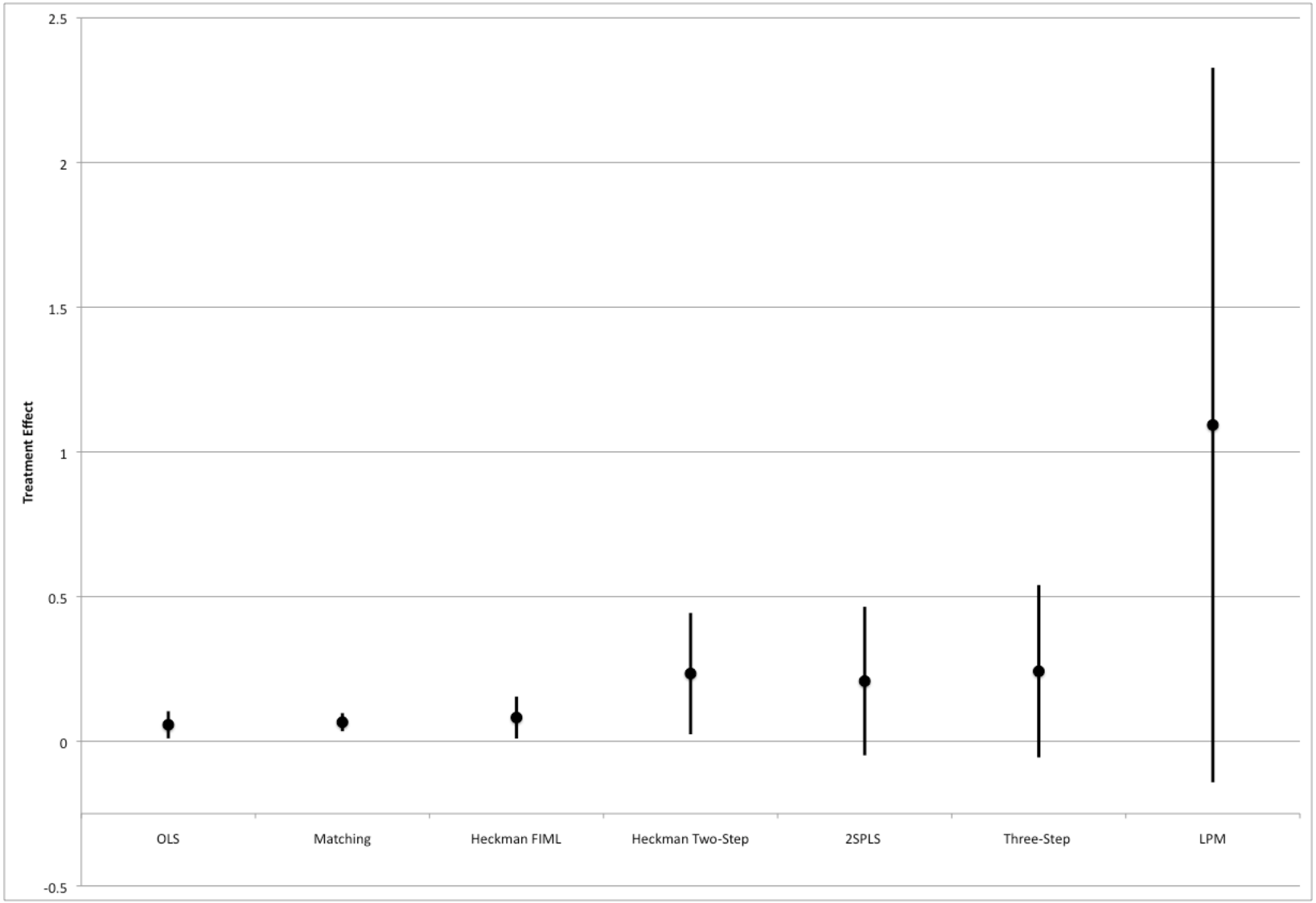


Table1. Bias of Treatment Effect Estimate

				Normal			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	-0.002	-0.154	-0.448	-0.748	-1.049	-1.348	-1.422
LPM	0.080	0.090	0.091	0.080	0.086	0.098	0.100
2SPLS	-0.002	0.013	0.009	-0.002	-0.002	0.016	0.022
Three-Step	-0.002	0.012	0.010	-0.003	-0.001	0.016	0.021
Heckman, Two-Step	-0.004	0.010	0.008	-0.004	-0.001	0.017	0.019
Heckman, FIML	-0.005	0.010	0.004	-0.006	-0.011	-0.002	0.003
				Normal			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	-0.002	-0.167	-0.487	-0.809	-1.135	-1.459	-2.539
LPM	0.065	0.116	0.202	0.160	0.259	0.289	0.388
2SPLS	-0.014	0.043	0.079	0.038	0.082	0.090	0.162
Three-Step	-0.028	0.048	0.087	0.061	0.169	0.126	0.188
Heckman, Two-Step	-0.030	0.023	0.040	-0.001	0.021	0.067	0.082
Heckman, FIML	-0.017	-0.012	-0.100	-0.135	-0.067	-0.007	0.000
				Logit			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	-0.003	-0.147	-0.430	-0.718	-1.007	-1.293	-1.364
LPM	0.107	0.117	0.117	0.106	0.111	0.122	0.123
2SPLS	-0.010	0.004	0.000	-0.011	-0.010	0.005	0.011
Three-Step	-0.001	0.012	0.009	-0.004	-0.001	0.014	0.018
Heckman, Two-Step	-0.003	0.012	0.014	0.006	0.015	0.035	0.037
Heckman, FIML	-0.004	0.012	0.008	0.009	0.055	0.113	0.099
				Uniform			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	-0.002	-0.165	-0.484	-0.807	-1.133	-1.456	-1.537
LPM	0.008	0.022	0.023	0.012	0.015	0.036	0.040
2SPLS	0.023	0.039	0.036	0.024	0.023	0.047	0.054
Three-Step	-0.002	0.014	0.012	0.000	0.000	0.024	0.031
Heckman, Two-Step	-0.004	0.005	-0.008	-0.035	-0.045	-0.036	-0.034
Heckman, FIML	-0.005	0.004	-0.009	-0.043	-0.141	-0.304	-0.308

Table 2. MSE of Treatment Effect Estimate

				Normal			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.005	0.028	0.205	0.562	1.102	1.817	2.023
LPM	0.051	0.050	0.055	0.053	0.052	0.052	0.055
2SPLS	0.049	0.047	0.052	0.052	0.053	0.052	0.051
Three-Step	0.049	0.046	0.052	0.051	0.052	0.050	0.051
Heckman, Two-Step	0.048	0.045	0.050	0.049	0.051	0.048	0.048
Heckman, FIML	0.049	0.045	0.046	0.037	0.021	0.007	0.004
				Normal			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.005	0.032	0.241	0.658	1.289	2.130	2.371
LPM	1.213	1.050	5.519	1.068	3.262	6.227	3.863
2SPLS	0.603	0.634	0.655	0.676	0.767	0.732	0.829
Three-Step	0.847	1.354	0.720	0.912	2.860	0.948	1.209
Heckman, Two-Step	0.479	0.480	0.502	0.477	0.472	0.452	0.434
Heckman, FIML	0.348	0.324	0.311	0.264	0.109	0.011	0.005
				Logit			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.005	0.026	0.189	0.518	1.015	1.673	1.860
LPM	0.049	0.049	0.053	0.050	0.050	0.051	0.053
2SPLS	0.044	0.041	0.046	0.045	0.046	0.045	0.044
Three-Step	0.043	0.040	0.045	0.044	0.045	0.044	0.043
Heckman, Two-Step	0.042	0.039	0.045	0.044	0.045	0.044	0.044
Heckman, FIML	0.043	0.039	0.040	0.035	0.025	0.019	0.014
				Uniform			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.004	0.031	0.238	0.653	1.285	2.121	2.362
LPM	0.070	0.068	0.072	0.071	0.070	0.067	0.074
2SPLS	0.067	0.066	0.071	0.072	0.073	0.073	0.078
Three-Step	0.070	0.067	0.073	0.074	0.074	0.072	0.078
Heckman, Two-Step	0.065	0.063	0.068	0.068	0.069	0.064	0.066
Heckman, FIML	0.066	0.063	0.061	0.045	0.039	0.101	0.102

Table 3. Coverage of Treatment Effect Estimate

				Normal			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.947	0.348	0	0	0	0	0
LPM	0.952	0.972	0.973	0.981	0.987	0.991	0.993
2SPLS	0.975	0.978	0.991	0.994	0.995	0.994	0.995
Three-Step	0.961	0.957	0.946	0.941	0.946	0.961	0.957
Heckman, Two-Step	0.959	0.957	0.954	0.938	0.944	0.96	0.957
Heckman, FIML	0.956	0.95	0.945	0.934	0.956	0.957	0.93
				Normal			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.944	0.301	0	0	0	0	0
LPM	0.955	0.960	0.973	0.969	0.977	0.978	0.974
2SPLS	0.966	0.965	0.980	0.980	0.982	0.981	0.986
Three-Step	0.987	0.991	0.979	0.963	0.936	0.943	0.948
Heckman, Two-Step	0.981	0.980	0.976	0.955	0.947	0.950	0.943
Heckman, FIML	0.799	0.821	0.837	0.858	0.937	0.948	0.943
				Logit			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.939	0.402	0	0	0	0	0
LPM	0.939	0.961	0.965	0.972	0.979	0.984	0.984
2SPLS	0.975	0.979	0.991	0.992	0.993	0.996	0.996
Three-Step	0.96	0.956	0.953	0.943	0.944	0.957	0.961
Heckman, Two-Step	0.954	0.954	0.949	0.94	0.945	0.968	0.968
Heckman, FIML	0.954	0.948	0.947	0.931	0.905	0.67	0.585
				Uniform			
$\rho_{e1,e2}$	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**0.95**
OLS	0.949	0.28	0	0	0	0	0
LPM	0.979	0.976	0.992	0.993	0.995	0.994	0.993
2SPLS	0.978	0.977	0.988	0.989	0.993	0.992	0.994
Three-Step	0.962	0.959	0.951	0.946	0.944	0.954	0.952
Heckman, Two-Step	0.963	0.954	0.949	0.94	0.934	0.935	0.932
Heckman, FIML	0.958	0.948	0.941	0.935	0.866	0.077	0.023

Table 4. Treatment Effect Estimates of Going Public

	Coefficient	*Standard Error*	*p value*	*p value of endogeneity test*
OLS	0.057	0.024	0.019	--
Matching	0.066	0.016	0.005	--
Heckman FIML	0.082	0.037	0.028	0.47
Heckman Two-Step	0.234	0.107	0.029	0.09
2SPLS	0.208	0.131	0.112	0.21
Three-Step	0.242	0.152	0.112	0.21
LPM	1.093	0.630	0.083	0.01
N=1124				

Table A1. Endogeneity as a Consequence of Omitted Variable Bias

	W Omitted				W Included			
	Outcome OLS	Outcome Heckman	Outcome 2-Stage	Treatment Selection (Probit)	Outcome OLS	Outcome Heckman	Outcome 2-Stage	Treatment Selection (Probit)
D	1.957 (.095)	0.678 (.217)	0.691 (.219)	–	0.985 (.077)	0.828 (.128)	0.800 (.130)	–
X	0.869 (.047)	1.135 (.064)	1.132 (.065)	0.736 (.054)	1.035 (.035)	1.067 (.040)	1.071 (.041)	1.084 (.079)
W	–	–	–	–	-1.018 (.035)	-1.049 (.040)	-1.055 (.041)	-1.087 (.077)
Z	–	–	–	0.740 (.055)	–	–	–	1.142 (.082)
Lambda	–	0.964 (.139)	–	0.740 (.055)	–	0.141 (.092)	–	1.142 (.082)
Constant	-0.421 (.063)	0.207 (.116)	0.200 (.117)	-0.023 (.046)	0.011 (.049)	0.087 (.069)	0.099 (.070)	-0.096 (.056)

Table A2. Complete Model Estimates for Going Public (Canes-Wrone 2001)

Outcome		**OLS**		**Heckman FIML**			**Heckman Two-Step**		
Variable	*Coef.*	*S.E.*	*p*	*Coef.*	*S.E.*	*p*	*Coef.*	*S.E.*	*p*
Public Appeal	0.057	0.024	0.019	0.082	0.037	0.028	0.234	0.107	0.029
Unified Government	0.024	0.023	0.298	0.026	0.023	0.262	0.037	0.025	0.134
Prior Media Salience	-0.006	0.012	0.636	-0.008	0.012	0.532	-0.019	0.015	0.189
Most Important Problem	-0.021	0.021	0.335	-0.024	0.022	0.267	-0.044	0.026	0.087
Priority	-0.022	0.016	0.169	-0.024	0.016	0.138	-0.034	0.018	0.053
Targeted Address	0.010	0.040	0.809	0.006	0.040	0.886	-0.018	0.044	0.679
Personal Popularity	0.017	0.015	0.268	0.017	0.015	0.272	0.015	0.016	0.333
Honeymoon	0.010	0.015	0.527	0.010	0.015	0.517	0.011	0.015	0.482
% Change GDP	-0.002	0.003	0.576	-0.002	0.003	0.580	-0.001	0.003	0.627
Kennedy	-0.039	0.029	0.175	-0.042	0.029	0.143	-0.062	0.032	0.056
Johnson	0.005	0.023	0.819	0.003	0.023	0.904	-0.013	0.026	0.626
Nixon	0.035	0.024	0.142	0.037	0.024	0.123	0.046	0.025	0.067
Ford	-0.009	0.032	0.776	-0.009	0.032	0.783	-0.007	0.033	0.841
Reagan	-0.035	0.022	0.109	-0.034	0.021	0.110	-0.032	0.022	0.145
Bush	-0.039	0.026	0.140	-0.038	0.026	0.153	-0.029	0.028	0.301
Clinton	-0.120	0.027	0.000	-0.120	0.026	0.000	-0.123	0.027	0.000
Constant	-0.078	0.021	0.000	-0.079	0.020	0.000	-0.086	0.021	0.000
First Stage									
Agency Size/Instrument				0.224	0.100	0.025	0.215	0.099	0.031
Unified Government				-1.464	0.527	0.005	-1.482	0.531	0.005
Prior Media Salience				0.172	0.113	0.127	0.177	0.113	0.119
Most Important Problem				0.979	0.204	0.000	0.972	0.205	0.000
Priority				0.388	0.166	0.019	0.395	0.167	0.018
Targeted Address				0.816	0.324	0.012	0.815	0.325	0.012
Personal Popularity				0.315	0.219	0.151	0.322	0.219	0.142
Honeymoon				-0.069	0.188	0.716	-0.073	0.188	0.697
% Change GDP				-0.032	0.038	0.396	-0.033	0.038	0.387
Kennedy				1.898	0.578	0.001	1.921	0.584	0.001
Johnson				1.803	0.591	0.002	1.828	0.597	0.002
Nixon				-0.823	0.361	0.022	-0.824	0.360	0.022
Ford				0.098	0.379	0.796	0.107	0.378	0.777
Reagan				0.064	0.261	0.807	0.073	0.261	0.780
Bush				-0.664	0.348	0.056	-0.656	0.347	0.059
Clinton				0.323	0.363	0.373	0.324	0.364	0.375
Constant				-1.951	0.251	0.000	-1.954	0.251	0.000
rho				-0.069		0.472	-0.477		0.089

Table A2 (Continued). Complete Model Estimates for Going Public (Canes-Wrone 2001)

Outcome		**2SPLS**			**Three-Step**			**LPM**	
Variable	*Coef.*	*S.E.*	*p*	*Coef.*	*S.E.*	*p*	*Coef.*	*S.E.*	*p*
Public Appeal	0.208	0.131	0.112	0.242	0.152	0.112	1.093	0.630	0.083
Unified Government	0.035	0.026	0.165	0.038	0.026	0.148	0.079	0.051	0.119
Prior Media Salience	-0.018	0.016	0.272	-0.020	0.017	0.239	-0.080	0.049	0.106
Most Important Problem	-0.040	0.027	0.146	-0.045	0.030	0.126	-0.122	0.071	0.089
Priority	-0.032	0.018	0.082	-0.035	0.019	0.071	-0.059	0.035	0.092
Targeted Address	-0.015	0.046	0.748	-0.020	0.047	0.681	-0.142	0.113	0.211
Personal Popularity	0.015	0.016	0.325	0.015	0.016	0.338	0.008	0.026	0.763
Honeymoon	0.011	0.015	0.470	0.011	0.015	0.481	0.022	0.026	0.391
% Change GDP	-0.001	0.003	0.614	-0.001	0.003	0.630	-0.001	0.005	0.855
Kennedy	-0.059	0.034	0.082	-0.063	0.035	0.074	-0.134	0.075	0.075
Johnson	-0.011	0.027	0.695	-0.013	0.028	0.634	-0.078	0.064	0.221
Nixon	0.044	0.025	0.085	0.047	0.026	0.074	0.080	0.048	0.095
Ford	-0.007	0.033	0.822	-0.006	0.033	0.844	-0.006	0.052	0.912
Reagan	-0.032	0.022	0.144	-0.032	0.022	0.148	-0.006	0.039	0.879
Bush	-0.029	0.028	0.306	-0.028	0.028	0.322	0.019	0.056	0.733
Clinton	-0.123	0.027	0.000	-0.123	0.027	0.000	-0.134	0.044	0.002
Constant	-0.085	0.022	0.000	-0.086	0.022	0.000	-0.132	0.046	0.005
First Stage									
Agency Size/Instrument	0.215	0.099	0.031	0.860	0.156	0.000	0.030	0.017	0.070
Unified Government	-1.482	0.531	0.005	-0.010	0.031	0.752	-0.058	0.020	0.003
Prior Media Salience	0.177	0.113	0.119	0.010	0.019	0.592	0.060	0.022	0.006
Most Important Problem	0.972	0.205	0.000	0.022	0.033	0.503	0.105	0.034	0.002
Priority	0.395	0.167	0.018	0.012	0.022	0.580	0.026	0.020	0.199
Targeted Address	0.815	0.325	0.012	0.020	0.055	0.722	0.134	0.077	0.082
Personal Popularity	0.322	0.219	0.142	0.002	0.019	0.936	0.010	0.012	0.388
Honeymoon	-0.073	0.188	0.697	0.001	0.018	0.951	-0.012	0.011	0.277
% Change GDP	-0.033	0.038	0.387	0.000	0.003	0.939	-0.001	0.002	0.751
Kennedy	1.921	0.584	0.001	0.016	0.040	0.693	0.094	0.035	0.007
Johnson	1.828	0.597	0.002	0.011	0.033	0.729	0.080	0.020	0.000
Nixon	-0.824	0.360	0.022	-0.010	0.031	0.733	-0.049	0.020	0.015
Ford	0.107	0.378	0.777	-0.004	0.039	0.923	-0.004	0.028	0.875
Reagan	0.073	0.261	0.780	-0.001	0.026	0.966	-0.026	0.021	0.232
Bush	-0.656	0.347	0.059	-0.003	0.034	0.923	-0.056	0.022	0.011
Clinton	0.324	0.364	0.375	-0.001	0.033	0.983	0.014	0.022	0.520
Constant	-1.954	0.251	0.000	0.006	0.026	0.830	0.048	0.020	0.020