Beyond LATE: A Simple Method for Estimation of the Average Treatment Effect with an Instrumental Variable

Peter M. Aronow* and Allison Carnegie

August 30, 2012
Abstract

Political scientists frequently use instrumental variables estimation to estimate the causal effect of an endogenous treatment variable. However, when the treatment effect is heterogeneous, this estimation strategy only recovers the Local Average Treatment Effect (LATE). The LATE is an average treatment effect for one subset of the population: units that receive treatment if and only if they are induced by an exogenous instrumental variable. Typically, researchers are interested in the average treatment effect (ATE) for the entire population of interest. In this paper, we highlight the important distinction between these two estimands and develop a simple and intuitive method for estimating the ATE even when treatment effects are heterogeneous. We apply our method to two published experiments in political science in which we demonstrate that the LATE can differ considerably from the ATE.

Keywords: Causal inference, compliance score, instrumental variables, local average treatment effect, average treatment effect
Introduction

Instrumental variables (IV) estimation, although commonly used in political science, is subject to a critical and often overlooked problem. IV estimation allows researchers to estimate the causal effect of an endogenous treatment variable by replacing the endogenous variable with an exogenous proxy variable known as an instrumental variable. However, although most researchers intend to estimate the average treatment effect (ATE) for their entire population, IV estimation only recovers the local average treatment effect (LATE), or the average treatment effect for the subset of the population that is influenced by the instrumental variable. When the effect of the endogenous variable is heterogeneous, interpretation of the estimate becomes complicated. Given that treatment effects are likely heterogeneous in most social science applications, problems of interpretation are quite common. To address these problems, we provide researchers with a new method for estimating the ATE in many studies that rely on IV estimation.

To understand the distinction between the LATE and ATE, consider, for example, a randomized experiment designed to assess the effect of election monitoring on incumbent vote share (and other dependent variables) (Hyde 2010). Villages were randomly assigned to receive election monitoring; however, some villages assigned to receive monitors failed to receive them and some villages not assigned to receive monitors received them regardless. To address this issue of noncompliance, the treatment assignment may be used as an instrumental variable for treatment receipt. This estimation strategy, as conducted with two-stage least squares, consistently estimates the LATE: the effect of election monitoring among villages that would receive monitors if and only if they were assigned to receive them. In contrast, the ATE would represent the effect of monitoring on all villages, regardless of whether or not they would comply with their treatment assignments.
When is the fact that IV estimation only recovers the LATE a problem for researchers? The answer to this question depends on the research agenda. For example, if the researcher is interested in the effect of election monitoring among villages that would comply with the treatment assignment, then the LATE is the desired estimand. However, if the researcher is interested in the effect of monitoring among all villages, the LATE will not represent the quantity of interest. While the LATE may have some utility for, e.g., cost considerations, the ATE is generally considered a more theoretically important estimand (Imbens 2009).

The goal of this paper is not to dissuade researchers from using IV estimation in experimental analysis. Indeed, IV estimation is a valuable tool, as it allows researchers to overcome common problems of selection bias and unobserved heterogeneity to consistently estimate an internally valid causal quantity for a subset of the population. But it is important for researchers to recognize the distinction between the LATE and the ATE. Interpretation of the LATE as a causal quantity of interest is clearly a major concern for any researcher who contends with noncompliance in experimental methods, but one that is rarely addressed in applied work in political science. In fact, in a review of 34 empirical articles that employ IV estimation from 2004-2009 in the American Political Science Review and the American Journal of Political Science, only two (6%) mention that the causal effect being estimated is the LATE.

This paper presents a solution to these problems of interpretation by developing a simple and intuitive method for estimating the ATE. Our method is easily implemented using the logic of two procedures already familiar to political scientists: inverse probability weighting (IPW) for sample correction and IV estimation. We begin by reviewing the assumptions and limitations of IV estimation, first from a structural equation modeling (SEM) perspective and then using the more flexible potential outcomes interpretation
(Angrist, Imbens and Rubin 1996). We next explain our method for recovering the ATE. We provide a novel maximum likelihood estimator to use covariates to predict compliance scores for units, even in the presence of two-sided noncompliance. We show researchers how to use the estimated compliance scores to reweight units using a method that we refer to as inverse compliance score weighting (ICSW), which is analogous to IPW. Used in conjunction with a standard IV estimator, ICSW allows researchers to estimate the ATE in studies with binary assignment and treatment variables. We then apply our method to two published papers in political science and find substantively meaningful differences between the LATE and ATE.

**Instrumental Variables Estimation**

Political scientists frequently contend with problems of endogeneity, which occur when a variable of interest is systematically related to unobserved causes of the outcome variable. Endogeneity can arise for a variety of reasons, including selection bias, simultaneity, or measurement error. A popular method for dealing with endogeneity is IV estimation, which can be understood through two frameworks: the SEM framework and the potential outcomes framework. The SEM framework is most familiar to political scientists, and so we begin with it. However, IV estimation in the context of this framework introduces the rigid – and often implausible – assumption of constant treatment effects. As we proceed, this assumption will be relaxed, but at the cost of redefining the causal estimand.

**IV in an SEM Framework**

The SEM interpretation of IV estimation posits a linear, additive relationship between an outcome variable \( (Y_i) \), an endogenous treatment variable \( (D_i) \), covariates
\((Q_{1i}, Q_{2i}, \ldots, Q_{Ki})\), and an unobserved error term \((u_i)\) for each unit \(i\), such that

\[
Y_i = \beta_0 + \beta_1 D_i + \lambda_1 Q_{1i} + \lambda_2 Q_{2i} + \ldots + \lambda_K Q_{Ki} + u_i
\] (1)

We are able to consistently estimate \(\beta_1\) using ordinary least squares (OLS) regression when \(D_i\) is asymptotically uncorrelated with the error term \((\text{plim} \frac{1}{N} \Sigma D_i u_i = 0)\). This requirement is violated, however, when \(D_i\) is endogenous.

A solution to the endogeneity problem is to find an instrument \(Z_i\) that is only predictive of \(Y_i\) through its impact on \(D_i\). In other words, \(Z_i\) affects \(D_i\), but is excluded from equation (1). We can then use instrumental variables regression, which is based on a two equation model. The first equation, or the “first stage,” is a regression of the endogenous variable on the instrument and covariates:

\[
D_i = \gamma_0 + \gamma_1 Z_i + \delta_1 Q_{1i} + \delta_2 Q_{2i} + \ldots + \delta_K Q_{Ki} + e_i
\] (2)

where \(e_i\) is a second unobserved error term. The coefficients from this first stage regression generate predicted values of \(D_i\), purging \(D_i\) of endogeneity. We can substitute the predicted values of \(D_i\) back into equation (1), regressing \(Y_i\) on the predicted values of \(D_i\) as well as the covariates. This two-stage substitution process is what gives the regression the name two-stage least squares (2SLS).

Given the model in equations (1) and (2), the SEM instrumental variables framework relies on two main assumptions for consistency. First, the correlation between \(Z_i\) and \(u_i\) must approach zero as \(N\) grows. This assumption cannot be tested empirically and must be justified based on theoretical knowledge. Second, the covariance between \(Z_i\) and \(D_i\) must converge to a nonzero quantity as \(N\) grows. The second assumption can be tested empirically by examining the correlation between \(Z_i\) and \(D_i\) (Sovey and Green 2011).
Note, however, that while the endogenous variable and covariates in equation (1) are indexed by the subscript $i$, their coefficients (the $\beta$s and $\lambda$s) are not. In other words, the SEM framework assumes constant treatment effects, so that the effect of the treatment is identical for each unit $i$. The constant effects assumption is very strong and can rarely be justified in social science applications. For example, consider the implications of the constant effects assumption in the context of an experiment designed to assess the effects of negative campaign advertising on political attitudes (Arceneaux and Nickerson 2010). Constant effects implies that a negative advertisement affects attitudes for all subjects, from those who have strong prior political opinions to those who are indifferent, by exactly the same amount. We can avoid making the constant treatment effects assumption using the potential outcomes framework developed by Angrist, Imbens and Rubin (1996), which shows that the IV estimator can identify the LATE even when treatment effects are heterogeneous. In the following section, we introduce the potential outcomes notation and discuss the implications of identifying the LATE.

### IV in a Potential Outcomes Framework

We now move to the Neyman-Rubin potential outcomes framework, focusing on the case of a binary instrumental variable (or treatment assignment) $Z_i$ and a binary endogenous treatment variable (or treatment received) $D_i$. For unit $i$, let $Y_{0i}$ be the outcome if untreated, $D_i = 0$, and let $Y_{1i}$ be the outcome if treated, $D_i = 1$. The treatment effect for a given unit $i$ is the difference between this unit’s outcomes in both possible states of the world, $Y_{1i} - Y_{0i}$. Similarly, we may define $D_{0i}$ as the treatment condition of unit $i$ when assigned to control, $Z_i = 0$, and $D_{1i}$ as the treatment condition of unit $i$ when assigned to treatment, $Z_i = 1$. We will show that the causal estimand recovered by IV estimation will not generally be the ATE, $E(Y_{1i} - Y_{0i})$, but will instead be the LATE,
$$E(Y_{1i} - Y_{0i} | D_{1i} > D_{0i})$$

Following Angrist, Imbens and Rubin (1996), the population may be divided into four groups: always-takers, never-takers, compliers, and defiers. Always-takers are units that receive treatment regardless of whether they are assigned to treatment or control, so that $D_{0i} = 1$ and $D_{1i} = 1$. Conversely, never-takers do not receive treatment regardless of their treatment assignment, so that $D_{0i} = 0$ and $D_{1i} = 0$. Compliers receive treatment if assigned to treatment and do not receive treatment if assigned to control, so that $D_{0i} = 0$ and $D_{1i} = 1$. (or $D_{1i} > D_{0i}$). Defiers receive treatment if assigned to control and do not receive treatment if assigned to treatment, so that $D_{0i} = 1$ and $D_{1i} = 0$ (or $D_{0i} > D_{1i}$). In the example of a randomized clinical trial to assess the causal effect of taking a medication, always-takers will take the medication regardless of which group they are assigned to, never-takers will not take the medication regardless of which group they are assigned to, compliers will take the medication if and only if assigned to the treatment group, and defiers will take the medication if and only if assigned to the control group. Although these latent subgroups may be interpreted deterministically, they also have a probabilistic interpretation that we employ. Unit $i$’s membership in each of these subgroups may be viewed as a fundamentally stochastic quantity. If the study is interpreted as a sample from a larger population of potential studies, then, e.g., unit $i$ could be an always-taker in one realization of the study and a complier in a different realization of the study.

Using an IV estimator, we may estimate the LATE using five assumptions, as explicated by Angrist, Imbens and Rubin (1996). First, we assume that the exclusion restriction is valid, or that $Z_i$ only affects $Y_i$ through $D_i$. Second, we assume that $Z_i$ to some degree predicts $D_i$. Third, we invoke the non-interference assumption, which states that the potential outcomes $D_{0i}, D_{1i}, Y_{0i},$ and $Y_{1i}$ are invariant with respect to the particular arrangement of the treatment vectors $Z$ and $D$. Fourth, we assume that the population
contains no defiers, i.e., \( \Pr(D_{0i} > D_{1i}) = 0 \). Fifth, we assume that \( Z_i \) is randomly assigned. As no parametric assumptions are necessary, these assumptions are considerably weaker than (and subsumed by) those of the SEM framework.

Angrist, Imbens and Rubin (1996) demonstrate that these five assumptions imply that the “intention-to-treat” effect of \( Z_i \) on \( Y_i \) (the average effect of treatment assignment on the outcome) divided by the intention-to-treat effect of \( Z_i \) on \( D_i \) (the average effect of treatment assignment on treatment received) is equal to the average causal treatment effect for compliers:

\[
E(Y_{1i} - Y_{0i} | D_{1i} > D_{0i}) = \frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)}.
\]

The LATE, \( E(Y_{1i} - Y_{0i} | D_{1i} > D_{0i}) \), can therefore be estimated with

\[
\hat{\tau}_{LATE} = \frac{(\sum_{i=1}^{N} w_i Z_i Y_i)/(\sum_{i=1}^{N} w_i Z_i) - (\sum_{i=1}^{N} w_i (1-Z_i) Y_i)/(\sum_{i=1}^{N} w_i (1-Z_i))}{(\sum_{i=1}^{N} w_i Z_i D_i)/(\sum_{i=1}^{N} w_i Z_i) - (\sum_{i=1}^{N} w_i (1-Z_i) D_i)/(\sum_{i=1}^{N} w_i (1-Z_i))}
\]

or, equivalently, the bivariate 2SLS estimator with weights \( w_i \).\(^1\) This estimator is consistent but not unbiased; it is subject to finite sample bias due to the fact that it is a ratio estimator. When \( Z_i \) is randomly assigned (and therefore unrelated to covariates), the addition of covariates to the 2SLS model will also yield consistent estimation of the LATE (Angrist and Pischke 2009).\(^2\) Intuitively, 2SLS simply rescales the intention-to-treat ef-

---

\(^1\)\( w_i \) represents any appropriate weights applied to the data. These weights facilitate correction for missingness (Samii 2011), unequal probabilities of assignment to treatment (Humphreys 2009; Glynn and Quinn 2010), and extrapolation to arbitrary populations (Horvitz and Thompson 1952). For one particularly useful example, if a researcher wished to only draw inferences about units with a particular covariate profile, \( w_i \) could be set to zero for all units that do not have this covariate profile.

\(^2\) We provide a simple intuition for why, under random assignment of \( Z_i \), 2SLS with covariates is consistent for the LATE. The weighted least squares regression of \( Y_i \) on \( Z_i \) and covariates is consistent for the numerator of equation (3) and the weighted least squares regression of \( D_i \) on \( Z_i \) and covariates is consistent for the denominator of equation (3). The 2SLS estimator of the effect of \( D_i \) will be the ratio of the two least squares estimators and, by Slutsky’s Theorem, will be consistent for the LATE.
fect of $Z_i$ on $Y_i$, acknowledging that only compliers will respond to changes in $Z_i$.

Because IV estimation provides asymptotically unbiased estimates of the LATE even when treatment effects are heterogeneous, IV estimation has become the de facto standard for causal inference in studies with noncompliance in the social sciences. We have demonstrated that IV recovers the LATE under weak assumptions. However, as discussed in the introduction, researchers are often interested in the ATE rather than the LATE. In the next section, we present a method that allows for estimation of the ATE.

**From LATE to ATE**

To this point, we have demonstrated that IV estimation can be a mixed blessing. On the one hand, IV estimation recovers the LATE under weak assumptions. On the other hand, the LATE is not always the causal estimand of interest. To address this problem, we provide a method to estimate the ATE using the intuition behind IPW, a commonly used method to estimate population-level quantities from nonrepresentative samples. IPW weights units by the inverse of the probability of sample inclusion in order to estimate population-level quantities. ICSW follows the same logic, but weights units by the inverse of the compliance score, or a unit’s probability of being a complier. We next introduce the compliance score and present a new maximum likelihood estimator to estimate compliance scores even in the presence of two-sided non-compliance.

**The Compliance Score**

The compliance score is a latent pretreatment covariate that represents unit $i$’s probability of complying with its treatment assignment. Recall that we have defined unit $i$’s membership in each of the three latent subgroups (compliers, always-takers and never-takers)
as a fundamentally stochastic quantity. We can understand the compliance score in terms of these subgroups: the compliance score is the probability that unit \(i\) will belong to the complier subgroup. Formally, the compliance score, \(P_{Ci}\), is equal to \(\Pr(D_{1i} > D_{0i})\).

Although we cannot observe the compliance score, we can use observed pretreatment covariates to estimate the probability that certain “types” of units will comply. Consider a randomized experiment wherein a subject’s compliance with treatment assignment depends solely on his or her gender. If we estimate that 50 percent of men in the sample comply (by estimating the intention-to-treat effect of \(Z_i\) on \(D_i\) for men), we would estimate that each man has a 50 percent chance of compliance. In practice, we can use a slew of covariates \(X_i\) to predict compliance for each unit. When a fully saturated model is used, i.e., dummy variables for every value and interaction of all covariates are included in \(X_i\), the set of dummy variables completely represents each type of unit.

More formally, following the above nomenclature, the probability of compliance for unit \(i\) is \(\mathbb{E}(D_i|Z_i = 1) - \mathbb{E}(D_i|Z_i = 0)\). Given the assumptions outlined in the previous section, we know that the compliance score (and all other pretreatment covariates) is asymptotically orthogonal to \(Z\). In conjunction with a known covariate profile, the compliance score for a given unit is simple to estimate under one-sided non-compliance (when no units assigned to control will ever receive treatment), as \(\mathbb{E}(D_i|Z_i = 1) - \mathbb{E}(D_i|Z_i = 0) = \mathbb{E}(D_i|Z_i = 1)\). When there are no always-takers (or defiers), all units assigned to treatment who take the treatment are compliers. Therefore, for all units in the treatment group, we can run any binomial regression of \(D\) on \(X\) to estimate the predicted probabilities of compliance.

We present a new maximum likelihood estimation technique (similar to that of Yau and Little 2001) that generalizes existing methods of compliance score estimation to the
case of two-sided non-compliance. As derived in appendix A, we can estimate the compliance score with a two step procedure. First, compute

$$\{\hat{\theta}_{A,C}, \hat{\theta}_{A|A,C}\} = \arg\max_{\theta_{A,C}, \theta_{A|A,C}} \left[ L(\theta_{A,C}, \theta_{A|A,C}|D, Z) \right] = \arg\max_{\theta_{A,C}, \theta_{A|A,C}} [\prod_{i=1}^{N} \left( (F(\theta_{A,C}X_i)(1 - F(\theta_{A|A,C}X_i))Z_i + F(\theta_{A,C}X_i)F(\theta_{A|A,C}X_i))^{D_i} ight. \\
\left. (1 - F(\theta_{A,C}X_i)(1 - F(\theta_{A|A,C}X_i))Z_i - F(\theta_{A,C}X_i)F(\theta_{A|A,C}X_i))^{1-D_i} \right)^{w_i}) \right], \quad (5)$$

where $\theta_{A|A,C}$ and $\theta_{A,C}$ are coefficients to be estimated, $X_i$ is a sufficient vector of predictive covariates for unit $i$, $w_i$ is the weight assigned to unit $i$ and $F$ is an arbitrary cumulative distribution function (CDF). For the purposes of this paper, $F(\cdot) = \Phi(\cdot)$, the Normal CDF, although the choice of CDF is typically immaterial (in the same fashion as is logit/probit estimation). After estimating $\theta_{A,C}$ and $\theta_{A|A,C}$, we may estimate the compliance score for unit $i$,

$$\hat{P}_{Ci} = \hat{P}(D_{1i} > D_{0i}) = F(\hat{\theta}_{A,C}X_i)(1 - F(\hat{\theta}_{A|A,C}X_i)).$$

---

3We have not seen an estimator of this form in the literature, as it proposes a sensible nested structure for compliance that nevertheless reduces to a binomial ML under one-sided noncompliance and strata means under saturation. For examples of the use of the compliance score under one-sided non-compliance in biostatistics, see Follmann (2000), Joffe and Brensinger (2003), Joffe, Ten Have and Brensinger (2003) and Roy, Hogan and Marcus (2008). For Bayesian estimators of the compliance score under two-sided non-compliance, see Imbens and Rubin (1997) and Hirano et al. (2000); however, these methods require specification of a parametric model for the joint distribution of treatment assignment, compliance and outcomes. In contrast, our method requires only specification of a parametric model for the first stage, and nonparametric extensions are clearly possible via NPMLE.

4As suggested above, in the case of one-sided non-compliance, such that the probability of being an always-taker reduces to zero, the likelihood function reduces to:

$$L(\theta_C|D, Z) = \prod_{i=1}^{N} w_i ((F(\theta_CX_i)Z_i)^{D_i} + (1 - F(\theta_CX_i)Z_i)^{1-D_i}) \quad (1 - F(\theta_CX_i)(1 - F(\theta_CX_i))^{1-D_i} \right)^{w_i})$$

where $\theta_{A,C}$ is now $\theta_C$ since there are no always-takers. This likelihood function is a familiar one: it is the standard likelihood function for binomial regression, if only applied to units in the treatment condition. With one-sided non-compliance and $F(\cdot) = \Phi(\cdot)$, our maximum likelihood estimator therefore reduces to the probit estimator as applied only to units with $Z_i = 1$. 

10
Inverse Compliance Score Weighting

How does the compliance score help solve the problem of reliance on the LATE even when treatment effects are heterogeneous? Recall that the basic problem addressed by this paper is that IV estimation only generates estimates of the average treatment effect for compliers. We posit a new reweighting method that we refer to as inverse compliance score weighting. ICSW follows the logic of inverse probability weighting: if a type of unit is disproportionately sampled, we may reweight the sample to reflect the distribution of types in the population. The complier population may be interpreted as a non-random subsample of the population of interest. We use the inverse of the compliance score, or probability of being included in the sample of compliers, to reweight the entire sample such that the covariate distribution in the complier population is identical to the covariate distribution of the population of interest.

We provide a simple example to illustrate how the method works. Imagine an experiment where gender is the only determinant of compliance, such that 75% of males comply with their treatment assignments but only 10% of females comply with their treatment assignments. Suppose that the treatment effect for males is 0, but the treatment effect for females is 1. If males and females are represented equally in the population, then the ATE is $0.5 \times 0 + 0.5 \times 1 = 0.5$. The problem, of course, is that we cannot directly estimate the ATE due to noncompliance. IV estimation recovers the LATE, which we compute using a weighted average, $0.5 \times 0.75 \times 0 + 0.5 \times 0.10 \times 1 = 0.12$. In this example, the LATE is much lower than the ATE because females are underrepresented among compliers. Recovering the ATE with noncompliance entails weighting units by the inverse of the compliance score,

---

5In work developed contemporaneously with our paper, Angrist and Fernandez-Val (2010) present a similar method to reweight covariate-specific LATEs to target populations. Also, for an integrative framework for recovering the ATE along with other causal quantities, see Esterling, Neblo and Lazer (2011). Note, however, that the latter method is not robust to two-sided noncompliance. Additionally, see Frangakis and Rubin (1999) for a similar approach involving missing data.
then performing IV estimation. Again computing a weighted average, the estimand recovered by this procedure is the ATE: 

\[
0.5 \times 0.75 \times \frac{1}{2 \times 2} + 0.5 \times 0.10 \times \frac{1}{10} - 0.5 \times 0.10 \times \frac{1}{10} + 0.5 \times 0.75 \times \frac{1}{2 \times 2} = 0.5.
\]

We develop this intuition formally, first providing a general expression for the estimator, then detailing the assumptions necessary for consistency.

**ICSW Estimator**

Following from our use of the potential outcomes framework, the estimation procedure presented here requires the instrument \( Z_i \) and endogenous variable \( D_i \) to be binary. This requirement is satisfied in many empirical studies, including numerous randomized experiments with non-compliance.\(^6\) In the bivariate case (when there are no covariates), the estimator is simple. Define \( \hat{w}_{Ci} = 1/\hat{P}_{Ci} \). The ICSW estimator, \( \hat{\tau}_{ATE} = \)

\[
\frac{\sum_{i=1}^{N} w_i \hat{w}_{Ci} Z_i Y_i}{\sum_{i=1}^{N} w_i \hat{w}_{Ci} Z_i} - \frac{\sum_{i=1}^{N} w_i \hat{w}_{Ci} (1 - Z_i) Y_i}{\sum_{i=1}^{N} w_i \hat{w}_{Ci} (1 - Z_i)}.
\]

An equivalent estimator may be derived using weighted 2SLS (with weights \( w_i \hat{w}_{Ci} \)). As with the simple 2SLS case, covariates may also be included to reduce sampling variability without any consequence for the asymptotic bias of the estimator.

Like all estimators that rely on reweighting, this estimator is typically more variable than its unweighted counterpart, 2SLS, and is potentially subject to greater finite sample bias. The ICSW estimator, like 2SLS, has increased finite sample bias when compliance rates are low, and this problem is exaggerated when some units have very low compliance scores.\(^7\) In order to estimate confidence intervals, we recommend bootstrapping the entire procedure.

---

\(^6\)The method can be extended to applications using continuous endogenous variables and multivalued (or multiple) instruments. As Angrist and Imbens (1995) demonstrate, with continuous endogenous variables, 2SLS is simply a weighted average of grouped data IV estimators. With multiple instruments, 2SLS is a weighted average (efficient under homoskedasticity) of the 2SLS estimator for each instrument. Although beyond the scope of this paper, under the assumptions outlined here, the ATE could be estimated for each instrument and a weighted average could be produced for the overall ATE.

\(^7\)In order to prevent very small compliance scores from forming in finite samples, we Winsorize our data.
The bootstrap can reflect the known sampling process, including strata and cluster sampling. The bootstrap also permits estimated confidence intervals for the difference between the LATE and ATE.

ICSW Assumptions

In order to recover the ATE using ICSW, two assumptions (in addition to the instrumental variables assumptions described above) are required. The assumptions are the standard assumptions associated with IPW estimation, but we describe below the particular implications for ICSW. In Appendix B, we prove that ICSW is consistent.

(Assumption 1) Latent ignorability, conditional on a covariate profile, of compliance with respect to heterogeneous treatment effects (Esterling, Neblo and Lazer 2011; Frangakis and Rubin 1999) holds. Before describing the implications of the assumption, we first articulate Assumption 1 formally. Define categorical variable $h \in H = (1, 2, 3, ..., M)$ as a “type” of unit; that is, a unit with a known, observed covariate profile. By definition, estimates. Define $c$ as the $1/N^\alpha$th quantile of $\hat{P}_{Ci}$, where $\alpha$ is a positive calibrating constant. We replace all $\hat{P}_{Ci} < c$ with $c$. We recommend (and use in this paper) $\alpha = 0.275$, which performed well in simulation studies presented in the Supporting Information. This procedure, Winsorizing, is performed because very low probabilities can introduce instability in the estimates resulting from inverse probability weighting (Elliott 2009), but Winsorizing may sometimes increase finite sample bias. Asymptotically, assuming that $\alpha$ remains fixed, the Winsorization has no impact on estimation, as $c^{-p}_0$. In fact, non-Winsorized estimates represent a special case of this Winsorization process whenever $\alpha \to \infty$. This parameterization also encompasses 2SLS whenever $\alpha = 0$.

For (Augmented) IPW-style estimators, analytic variance estimators presuming correct model specification may be highly sensitive to model misspecification and inaccurate in finite samples, whereas the bootstrap tends to have better coverage (Funk et al. 2011). In our own simulation studies, assuming $\alpha = 0.275$, we find that the bootstrap has conservative coverage for ICSW in small samples, and proper coverage in larger samples. These studies are presented in the Supporting Information.

This specification, which is equivalent to having a fully saturated compliance score model, is made without loss of generality, but simplifies the presentation of the assumptions underlying the estimation procedure. If the compliance score model is not saturated, the additional assumption that the compliance score model has the correct functional form is required. We examine the consequences of failure of this assumption with simulation studies in the Supporting Information, finding that this assumption tends to be unimportant.
\[ \sum_h \Pr(H = h) = 1. \]

Similar to the logic behind inverse probability weighting for sample designs (Horvitz and Thompson 1952), Assumption 1 entails that 
\[ \mathbb{E}(Y_1 - Y_0 \mid H = h) = \mathbb{E}(Y_1 - Y_0 \mid D_1 > D_0, H = h): \]
the expected value of the treatment effect conditional on the type of unit is equal to the expected value of the treatment effect conditional on the type of unit and compliance, so that revelation of compliance gives no information about the treatment effect. Put simply, the average treatment effect for all units of type \( h \) is equal to the average treatment effect for compliers of type \( h \).

Ignorability assumptions are necessary for a variety of common methods including propensity score methods, matching and even linear regression. However, Assumption 1 is weaker than the typical ignorability assumption, which requires that both 
\[ \mathbb{E}(Y_1 \mid H = h) = \mathbb{E}(Y_1 \mid D_1 > D_0, H = h) \]
and 
\[ \mathbb{E}(Y_0 \mid H = h) = \mathbb{E}(Y_0 \mid D_1 > D_0, H = h). \]
Assumption 1 is less restrictive than the typical ignorability assumption as there is no need to account for baseline selection bias, only the determinants of treatment effect heterogeneity. For example, consider a randomized clinical trial of an experimental medicine. Assume the medicine may affect subjects differently depending only on certain known underlying medical conditions. The researcher might not be able to determine or measure all of the causes of compliance that are related to baseline treatment outcomes, e.g., wealth or risk-taking, so the typical ignorability assumption would not be satisfied. However, Assumption 1 might still hold; only the preexisting conditions that predict both compliance and heterogeneity need to be included in the conditioning set. Sometimes, of course, there might be no systematic treatment effect heterogeneity with respect to compliance and Assumption 1 would hold regardless of the variables included in the conditioning set.

We can more formally describe sufficient conditions for and consequences of Assumption 1. If all predictors of compliance are known, then Assumption 1 will clearly hold. But what if only some (or even none) of the predictors of compliance are observed?
In the case where the treatment effect is a constant value, $\tau$, across all units, Assumption 1 always holds: $E(Y_1 - Y_0 \mid H = h) = E(Y_1 - Y_0 \mid D_1 > D_0, H = h) = \tau$. If we were to accept the SEM interpretation of IV, where treatment effects are constant, Assumption 1 would always be satisfied and both 2SLS and ICSW would be consistent. More generally, even if treatment effects are heterogeneous, but still statistically independent of compliance, Assumption 1 would necessarily be satisfied, and 2SLS and ICSW would be consistent.

But what if treatment effects are heterogeneous and we have failed to identify the sufficient conditioning set? As Assumption 1 no longer holds, the ICSW estimator is now biased, even asymptotically. However, we can recognize intuitively that ICSW will still recover an internally valid causal effect, not on the full population of interest, but for some population that has a covariate profile that more closely resembles the full population. Manipulating equations 14 and 16 from Appendix B, the asymptotic bias for ICSW under failure of Assumption 1 will be

$$
\sum_h \Pr(H = h) \left[ E(Y_1 - Y_0 \mid D_1 > D_0, H = h) - E(Y_1 - Y_0 \mid H = h) \right].
$$

Whether or not bias is reduced (relative to 2SLS) is a function of the particular data at hand; in general, we expect that bias will typically be reduced whenever the identified types predict both heterogeneous treatment effects and compliance. If the ATE is the parameter of interest, then 2SLS represents the special case of ICSW where no covariates are included in the compliance score estimation process, and there is thus only one identified type. While there are cases where adding covariates can increase bias under misspecification, we nevertheless achieve a more principled estimate by conditioning on all known covariates.

Although we cannot demonstrate that Assumption 1 holds empirically, we may test
whether the compliance score captures at least some of the true variation in compliance levels. We may wish to know if the covariates, taken together, are more prognostic of compliance than would be expected by random chance. Once we have an estimated compliance score distribution for a given population, we may employ a permutation test to assess the efficacy of the compliance score identification. To conduct a permutation test, the researcher randomly sorts a portion of the data in order to assess whether a given test statistic is more extreme than would be expected by chance. Under the null hypothesis that the covariates are independent of the compliance scores, the sum-of-squared-residuals (or SSR) for $D$ based on the compliance score and treatment assignment should be identical regardless of the particular arrangement of covariates.

Define $\Pr(D_i = 1|X) = F(\theta_{A,C} X_i)(1 - F(\theta_{A|A,C} X_i))Z_i + F(\theta_{A,C} X_i)F(\theta_{A|A,C} X_i)$. Formally, $H_0 : \sum_{i=1}^{N} (D_i - \hat{\Pr}(D_i = 1|X))^2 = \sum_{i=1}^{N} (D_i - \hat{\Pr}(D_i = 1|X'))^2$, where $X'$ has had its rows randomly permuted, perhaps conditionally on a known stratification. By substituting in the estimates of these quantities, we may test this hypothesis. A $p$-value is computed by calculating the proportion of permutations that yield an SSR that is greater than or equal to the observed SSR:

$$p = \Pr\left(\sum_{i=1}^{N} (D_i - \hat{\Pr}(D_i = 1|X))^2 \geq \sum_{i=1}^{N} (D_i - \hat{\Pr}(D_i = 1|X'))^2\right).$$

Rejection of the null does not imply that Assumption 1 holds, but failure to reject the null suggests that researchers should be cautious about drawing inferences about the ATE if treatment effect heterogeneity is suspected.

(Assumption 2) The compliance score for all units is strictly bounded $\in (0, 1]$ so that

---

10While any test statistic $t = f(D, \tilde{X}, \tilde{P}_C)$ could be used here, SSR provides an easy and intuitive statistic, reflecting overall classification error. Through simulation studies, we demonstrate the desirable properties of the SSR test statistic in the Supporting Information, although alternative statistics, including the likelihood (equivalently likelihood ratio) or sum of absolute residuals, perform equally well.
each unit has a positive probability of compliance. Note that IV estimation already requires the assumption of nonzero compliance in the full population, and we generalize this assumption to every unit. Assuming nonzero compliance has both philosophical and practical implications. If a unit can never be a complier, a counterfactual with which to generate an ATE may not exist. Practically, if a unit has zero probability of compliance, the weighting procedure will produce infinite weights asymptotically, leading to an undefined estimate. Conditional on Assumption 1’s validity, Assumption 2 is therefore verifiable asymptotically.

Applications

To demonstrate both the application of our method and the consequences of estimating the ATE rather than the LATE, we present two case studies using data from Green, Gerber and Nickerson (2003) and Albertson and Lawrence (2009). We select these two articles in particular because they allow us to display the use of our method in the intuitive case of one-sided non-compliance as well as the more complex case of two-sided non-compliance. Additionally, the studies have very different research questions and policy implications, allowing us to discuss the theoretical and practical implications of ICSW in more depth.

Green, Gerber and Nickerson (2003)

Green, Gerber and Nickerson (2003) present results from a large-scale \( N = 18,933 \) field experiment designed to assess the effect of canvassing on voter turnout. Registered voters in Bridgeport, Columbus, Detroit, Minneapolis, Raleigh, and St. Paul were randomly assigned to treatment or control, where the treatment was encouragement to vote
in advance of the November 6, 2001 local elections. Encouragement was delivered in the form of face-to-face contact with nonpartisan student and community organization members. Because some citizens who were assigned to the treatment group could not be contacted, the study features one-sided noncompliance. To get around this limitation, the authors use the attempt to administer the treatment as an instrument for the receipt of the treatment to consistently estimate the LATE. These results are presented in column (1) of Table 1. Among compliers, canvassing is estimated to increase turnout by 5.6 percentage points with a confidence interval of (2.2, 8.9).\footnote{The estimates presented differ slightly from those in the original paper because the original paper used fixed effects, which is inconsistent in the presence of heterogeneous probabilities of assignment to treatment (Humphreys 2009; Angrist and Pischke 2009). We instead use inverse probability weighting, which is consistent in this example.}

The authors rightly point out that conducting an experiment represents a trade-off. On the one hand, randomization solves two common problems with observational studies of the determinants of voter mobilization: campaigns may target the voters most likely to go to the polls, preventing researchers from recovering the true effect of the campaign, and voters may not recall correctly whether they were contacted, introducing measurement error. On the other hand, the authors also state that “the principal complication that arises in experimental studies of voter mobilization is that some citizens assigned to the treatment group cannot be reached” (Green, Gerber and Nickerson 2003, 1085), which means the researchers can only estimate the LATE. Whether this represents a problem depends on whether the LATE is the estimand of interest.

Theoretically, the LATE may or may not interest researchers. Perhaps researchers care most about the effect of the treatment on subjects who can be contacted. After all, why concern oneself with the effect on subjects who will not receive the treatment? However, which subjects can be contacted is an artifact of the way in which the intervention
is deployed. For example, if canvassers attempted to contact subjects multiple times, the contact rate would surely have been higher. There is little theoretical reason to care about subjects who could be contacted in the context of this particular experiment. This theoretical shortcoming of the LATE is compounded by the fact that the authors are responding to a broader literature on the determinants of voter mobilization, citing works such as Rosenstone and Hansen (1993), Verba, Schlozman and Brady (1995), and Putnam (2000). These earlier works focus on the determinants of turnout in the entire population, while Green, Gerber and Nickerson (2003) estimate the determinants of turnout in the complier population, the answer to a potentially different question. Green, Gerber and Nickerson (2003, 1086) state that a central goal of their study is “to better gauge the average treatment effect of canvassing,” but the LATE provides limited information about the ATE. For example, non-compliers may have different education levels, income, interest in politics, or other demographic characteristics than compliers, leading them to respond differently to contact. Ex ante, we have no way of knowing whether the ATE would be higher or lower than the LATE.

The LATE, however, may be of practical interest. In this study, campaigns operated under tight budget constraints. If policy-makers plan to run another campaign under similar budget constraints, they will likely care most about the LATE. However, they may also care about the ATE, since the ATE would tell them the effect of a campaign that was able to contact everyone. If the ATE were much larger than the LATE, policy-makers may decide to expend more resources on the campaign. Thus, since both the LATE and the ATE may be useful for policy-makers, we estimate both.

We begin by estimating compliance scores for the sample using the covariates listed in Table 1 and dummy variables for the six cities in the sample. Applying our permutation test for the identification of compliance scores, we obtain $p < 0.000$, indicating
that we have identified at least some portion of the true compliance score distribution (see the Supporting Information for a plot). Although we cannot ascribe any causal relationship between compliance and covariates, we note that compliers are more likely to have voted previously, be members of a major party, and identify as white (see the Supporting Information for a correlation matrix between the estimated compliance score and covariates). We now apply ICSW and present the results in column (2) of Table 1. Canvassing is estimated to increase turnout in the entire population by 6.7 percentage points with a confidence interval of (3.2, 10.3), whereas it is estimated to increase turnout among compliers by only 5.6 percentage points. The ATE is therefore estimated to be 19 percent larger than the LATE.\textsuperscript{12} The estimated ATE strengthens Green, Gerber and Nickerson (2003, 1094)’s conclusion that “mobilization campaigns have the potential to increase turnout substantially in local elections,” and addresses the same population as the literature that the experiment is responding to.

**Albertson and Lawrence (2009)**

Albertson and Lawrence (2009) present findings from an experiment ($N = 507$) in which survey respondents in Orange County, California were randomly assigned to receive encouragement to view a Fox debate on affirmative action, which would take place on the eve of the 1996 presidential election. Shortly after the election, these respondents were re-interviewed. The post-election questionnaire asked respondents whether they viewed the debate, whether they supported a California proposition (209) to eliminate affirmative action (coded 1 if they supported it and 0 if not) and how informed they felt about the proposition (coded on a scale from 1-4 from least to most informed). The authors use a standard instrumental variable design to address the fact that some who were not assigned

\textsuperscript{12}The difference in coefficients, $\hat{\tau}_{ATE} - \hat{\tau}_{LATE}$, has a 95% confidence interval of (-0.2, 2.3) and a 90% confidence interval of (0.0, 2.2).
to treatment reported viewing the debate and some who were assigned to treatment did not report viewing the debate. This two-sided noncompliance was nontrivial: 55% of subjects assigned to watch the debate did not report viewing the debate, and 4% of subjects who were not assigned to watch the debate reported viewing. Albertson and Lawrence’s IV regression results show a positive relationship between program viewing and feeling more informed about the issue and a statistically insignificant, negative relationship between program viewing and support for the proposition among compliers. Albertson and Lawrence’s original findings are presented in columns (1) and (3) of Table 2.\textsuperscript{13}

Is the ATE or the LATE a more appropriate estimand in this case? Our estimand of interest depends on the research question. Albertson and Lawrence (2009, 276) state that the question they seek to answer is whether “civic-minded television ha[s] a lasting impact on those who watch,” which responds to an existing literature that “supports the view that television can be expected to inform viewers, make issues more salient, change viewers’ attitudes and possibly even affect their behavior.” Thus, the subject population of interest is defined to be viewers, but leaves ambiguous which viewers are of interest. It is unclear whether the authors are interested in the way the program would affect all people (ATE), the subjects who would elect to watch the program only if encouraged to do so (LATE), or some other population. It is possible, for example, that the LATE is the estimand of interest from a policy perspective, since a policy-maker may wish to know the effect of airing a debate after encouraging individuals to watch through advertising, but that the ATE is of theoretical interest because it addresses the potential impact of media on the population as a whole. Both the LATE and the ATE are therefore informative.

\textsuperscript{13}Note that our replication of their results differs slightly from their original results due to the fact that we use IPW with all covariates as well as the treatment assignment indicator to address missing values in the dependent variable. This IPW step is also included in the bootstrap procedure. As Albertson and Lawrence note, missing values on the outcome variable appear to be missing-at-random, at least with respect to the treatment indicator.
We first estimate compliance scores for the sample using the eight covariates used by Albertson and Lawrence (see Table 2 for a description of each of the covariates). The permutation test provides a p-value of 0.056, suggesting that the covariates are predictive of compliance (see the Supporting Information for a plot). The compliance score most strongly correlates with political interest, newspaper reading, and education (see the Supporting Information for a full correlation matrix). Recall that compliers are substantively unusual: compliers would not ordinarily watch the program but watch only because they were induced by the treatment assignment.

We now estimate the ATE using ICSW; our estimates of the ATE are presented in columns (2) and (4) of Table 2. Although Albertson and Lawrence (2009) find that compliers are 0.28 points more informed after viewing the program, we find that viewers in the entire subject population are 0.40 points more informed after viewing. We therefore estimate that the ATE is approximately 45% larger than the LATE. If the ATE were the real parameter of interest in this study, using the LATE to approximate it would lead to an underestimate of the ATE. Turning to our analysis of the effect of program viewing on support for the measure, we see that Albertson and Lawrence (2009)’s 2SLS estimate and our ICSW estimate are nearly identical at -0.07 points and -0.05 points, respectively. We therefore find that the effect of viewing the debate on support for compliers is very similar to the effect for the overall population, suggesting two possible interpretations. Since we cannot reject the null hypothesis of no treatment effect for either the LATE or the ATE, we may suspect that there is no effect on attitudes resulting from the treatment. Alternatively, if we believe there is a small treatment effect, this finding suggests that the effect on opinion is relatively homogeneous with respect to the population of interest.14

14 The difference in coefficients for knowledge, $\hat{\tau}_{ATE} - \hat{\tau}_{LATE}$, has a 95% confidence interval of (-
Together, the analysis of these two measures using ICSW highlights the fact that, ex ante, it is unclear how close the LATE will be to the ATE. As the estimates of the knowledge effect highlight, the estimands may be meaningfully different; as the opinion effect shows, the parameters may be almost identical. When the ATE is the parameter of interest, the uncertainty associated with the LATE may be highly problematic for both theoretical and practical interpretation.

**Conclusion**

Although the ATE is often the true parameter of interest, scholars typically focus on the LATE because it is frequently the only available causal estimand. However, the LATE may not be representative of the treatment effect in the general population and reliance on the LATE may lead to substantive conclusions that are different from those suggested by the ATE. We have shown that virtually any researcher in the social sciences who wishes to use IV estimation must contend with reliance on the LATE, and that the problem can produce results that have little to no meaningful interpretation. To help solve this critical problem, we have provided a method to recover the ATE using only assumptions standard to instrumental variables estimators and inverse probability weighting for sample correction. Using this method, we have demonstrated that recovery of the ATE produces a meaningful difference for the conclusions of two published studies in political science. ICSW thus allows researchers to estimate the ATE, a causal estimand previously considered out of reach, in a vast array of applications throughout the social sciences.

---

0.02,0.90) and a 90% confidence interval of (0.00,0.52). The difference in coefficients for opinion has a 95% confidence interval of (-0.12,0.16) and a 90% confidence interval of (-0.08,0.12).
Appendix A: Derivation of Likelihood Function

As defined above, $D_i$ is an indicator variable for treatment received and $Z_i$ is an indicator variable for treatment assigned. Further, we define $X_i$ as a matrix representing the sufficient set of predictive covariates for unit $i$. For convenience, we make three easily relaxed parametric assumptions. The first assumption is that the probability of being an always-taker or a complier is a function of covariates with a known distribution.

$$P_{A,C,i} = \text{Pr}(D_{1i} > D_{0i} \cup D_{0i} = 1) = F(\theta_{A,C} X_i), \quad (6)$$

where $P_{A,C,i}$ is the probability that unit $i$ is either a complier or an always-taker, $\theta_{A,C}$ is a vector of coefficients to be estimated and $F(\cdot)$ is the cumulative distribution function (CDF) for an arbitrary distribution. As mentioned above, for the purposes of this paper, we will use a probit model, so $F(\cdot) = \Phi(\cdot)$, where $\Phi$ is the Normal CDF. However, other binomial regression models, including logit and generalized additive models (Hastie and Tibshirani 1990) could be used. Second, we similarly specify

$$P_{A|A,C,i} = \text{Pr}(D_{0i} = 1 | D_{1i} > D_{0i} \cup D_{0i} = 1) = F(\theta_{A|A,C} X_i), \quad (7)$$

where $P_{A|A,C,i}$ is the probability that unit $i$ is an always-taker conditional on it being either an always-taker or a complier, and $\theta_{A|A,C}$ are coefficients to be estimated. Therefore, we may define the compliance score as $P_{C,i} = \text{Pr}(D_{1i} > D_{0i})$. Since we know, by definition, that compliers receive treatment if and only if assigned to treatment and that always-takers always receive treatment,

$$\text{Pr}(D_i = 1) = \text{Pr}(D_{1i} > D_{0i}) Z_i + \text{Pr}(D_{0i} = 1) = P_{A,C,i}(1 - P_{A|A,C,i}) Z_i + P_{A,C,i} P_{A|A,C,i}. \quad (8)$$
This expression represents a fully specified model for $\Pr(D_i = 1)$, and this value is strictly bounded within $(0, 1)$ since $Z_i \in \{0, 1\}$ and $\Pr_{A,C,i} \in (0, 1)$. This boundedness along with the binary nature of $Z_i$ allows us to specify our third assumption: $D_i$ is Bernoulli distributed and units are independent with weight $w_i$. Using equation (6), the likelihood of the model for any unit $i$ may now be specified as:

$$L(P_{A|A,C,i}, P_{A,C,i} \mid D, Z) = (P_{A,C,i}(1 - P_{A|A,C,i})Z_i + P_{A,C,i}P_{A|A,C,i})^{D_i}$$

(9)

$$= (1 - P_{A,C,i}(1 - P_{A|A,C,i})Z_i - P_{A,C,i}P_{A|A,C,i})^{1 - D_i}.$$  

Combining equations 4, 5 and 7,

$$L(\theta_{A,C}, \theta_{A|A,C} \mid D, Z) =$$

$$\Pi_{i=1}^{N} ((F(\theta_{A,C}X_i)(1 - F(\theta_{A|A,C}X_i)Z_i + F(\theta_{A,C}X_i)F(\theta_{A|A,C}X_i))^{D_i}$$

$$(1 - F(\theta_{A,C}X_i)(1 - F(\theta_{A|A,C}X_i))Z_i - F(\theta_{A,C}X_i)F(\theta_{A|A,C}X_i))^{1 - D_i})^{w_i}.$$  

(10)

**Appendix B: Proof of ICSW Consistency**

In order to formally describe the process, we return to the derivation of the LATE. Note that, as above, our proof relies on the asymptotic qualities of each of the estimated quantities. Our proof strategy is to derive expressions for the inverse compliance score weighted numerator (average intention to treat effect, or ITT) and denominator (average probability of compliance) of equation (3), thus obtaining the asymptotic value of the IV estimator after ICSW. For clarity, we remove $i$ subscripts to denote all units subject to any conditional probabilities articulated. As in equation (3), $\text{ITT} = \Pr(D_1 > D_0)E(Y_1 - Y_0 \mid D_1 > D_0)$,
which we may express as a weighted sum of the conditional ITTs:

\[ ITT = \sum_h \Pr(D_1 > D_0 \mid H = h) \Pr(H = h) \mathbb{E}(Y_1 - Y_0 \mid D_1 > D_0, H = h). \quad (11) \]

With the equations above, we are now able to apply ICSW: for each \( h \in H \), we multiply by the weight: \( \frac{1}{\Pr(D_1 > D_0 \mid H = h)} \), or the inverse of the compliance score. We then divide by the average weight across all units (to normalize the weight). We define the average weight \( \bar{w}_c = \sum_h \frac{\Pr(H = h)}{\Pr(D_1 > D_0 \mid H = h)}. \)

We can write the weighted ITT as follows:

\[ ITT^w = \frac{1}{\bar{w}_c} \sum_h \Pr(D_1 > D_0 \mid H = h) \Pr(H = h) \mathbb{E}(Y_1 - Y_0 \mid D_1 > D_0 \mid H = h). \quad (12) \]

Since the proportion of compliers terms cancel out, equation (12) reduces to:

\[ ITT^w = \frac{1}{\bar{w}_c} \sum_h \Pr(H = h) \mathbb{E}(Y_1 - Y_0 \mid D_1 > D_0, H = h). \quad (13) \]

Applying Assumption 1, we can rewrite equation (13) as \( ITT^w = \frac{1}{\bar{w}_c} \sum_h \Pr(H = h) \mathbb{E}(Y_1 - Y_0 \mid H = h) \). By the law of total probability, the reweighted numerator of the IV estimator is

\[ ITT^w = \frac{1}{\bar{w}_c} \sum_h \Pr(H = h) \mathbb{E}(Y_1 - Y_0 \mid H = h) = \frac{1}{\bar{w}_c} \mathbb{E}(Y_1 - Y_0). \quad (14) \]

We can now reweight the denominator of the IV estimator by expanding the definition of the complier population, weighting, and simplifying:

\[ \Pr(D_1 > D_0) = \sum_h \Pr(D_1 > D_0 \mid H = h) \Pr(H = h) \quad (15) \]
\[
\Pr(D_1 > D_0) = \frac{1}{w_c} \sum_h \frac{\Pr(D_1 > D_0 \mid H = h) \Pr(H = h)}{\Pr(D_1 > D_0 \mid H = h)} = \frac{1}{w_c} \sum_h \Pr(H = h) = \frac{1}{w_c}.
\]

(16)

Dividing equation (14) by equation (16), we have the asymptotic value of the ICSW estimator,

\[
\frac{ITT^w}{\Pr(D_1 > D_0)} = \frac{1}{w_c} \frac{\mathbb{E}(Y_1 - Y_0)}{\Pr(D_1 > D_0)} = \mathbb{E}(Y_1 - Y_0),
\]

(17)

which is the ATE. By Slutsky’s Theorem, a consistent estimator of \(\frac{ITT^w}{\Pr(D_1 > D_0)}\) will be,

\[
\frac{\sum_{i=1}^N w_i \hat{w}_{Ci} Z_i Y_i}{\sum_{i=1}^N w_i \hat{w}_{Ci} Z_i} - \frac{\sum_{i=1}^N w_i \hat{w}_{Ci} (1 - Z_i) Y_i}{\sum_{i=1}^N w_i \hat{w}_{Ci} (1 - Z_i)}
\]

\[
\sum_{i=1}^N w_i \hat{w}_{Ci} Z_i D_i - \sum_{i=1}^N w_i \hat{w}_{Ci} (1 - Z_i) D_i
\]

\[
\sum_{i=1}^N w_i \hat{w}_{Ci} Z_i D_i - \sum_{i=1}^N w_i \hat{w}_{Ci} (1 - Z_i) D_i
\]

\[
\rightarrow_p \text{ITT}^w \text{ and }
\]

\[
\Pr(D_1 > D_0) \text{ if } \hat{w}_{Ci} \text{ is estimated consistently (as would be the case with a fully saturated compliance score model).}
\]

References


Yau, Linda H.Y. and Roderick J. Little. 2001. “Inference for the Complier-Average Causal Effect from Longitudinal Data Subject to Noncompliance and Missing Data, with Ap-
plication to a Job Training Assessment for the Unemployed.” Journal of the American Statistical Association 96.
### Tables

<table>
<thead>
<tr>
<th>Voter Turnout in 2001 Local Elections</th>
<th>LATE (2SLS)</th>
<th>ATE (ICSW)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Campaign Contact</strong></td>
<td><strong>5.6</strong></td>
<td><strong>6.7</strong></td>
</tr>
<tr>
<td></td>
<td><strong>(2.2, 8.9)</strong></td>
<td><strong>(3.2, 10.3)</strong></td>
</tr>
<tr>
<td>Number of Family Members</td>
<td>-0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>White</td>
<td>9.0</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td><strong>(3.9, 13.7)</strong></td>
<td><strong>(4.5, 13.9)</strong></td>
</tr>
<tr>
<td>Black</td>
<td>-1.7</td>
<td>-1.1</td>
</tr>
<tr>
<td></td>
<td><strong>(-7.7, 3.5)</strong></td>
<td><strong>(-6.7, 3.7)</strong></td>
</tr>
<tr>
<td>Voted in 2000 General Election</td>
<td>23.0</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td><strong>(21.9, 24.1)</strong></td>
<td><strong>(19.8, 22.2)</strong></td>
</tr>
<tr>
<td>Voted in 1999 General Election</td>
<td>30.5</td>
<td>27.2</td>
</tr>
<tr>
<td></td>
<td><strong>(28.5, 32.4)</strong></td>
<td><strong>(25.1, 29.4)</strong></td>
</tr>
<tr>
<td>Voted in 2001 Primary Election</td>
<td>48.7</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td><strong>(47.0, 50.5)</strong></td>
<td><strong>(49.2, 52.9)</strong></td>
</tr>
<tr>
<td>Age</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td><strong>(0.2, 0.2)</strong></td>
<td><strong>(0.1, 0.2)</strong></td>
</tr>
<tr>
<td>Age Missing</td>
<td>12.7</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td><strong>(-16.5, 44.1)</strong></td>
<td><strong>(-15.9, 50.3)</strong></td>
</tr>
<tr>
<td>Democrat</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td><strong>(0.6, 4.7)</strong></td>
<td><strong>(0.7, 4.8)</strong></td>
</tr>
<tr>
<td>Republican</td>
<td>3.8</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td><strong>(1.3, 6.2)</strong></td>
<td><strong>(0.8, 5.8)</strong></td>
</tr>
<tr>
<td>Independent</td>
<td>8.2</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td><strong>(5.4, 10.7)</strong></td>
<td><strong>(3.8, 8.8)</strong></td>
</tr>
<tr>
<td>City Controls</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1: 2SLS and ICSW estimates for Green, Gerber and Nickerson (2003). Bootstrap 95% confidence intervals in parentheses. Dependent variable is voter turnout in the 2001 general election in percentage points. Covariates include dummy variables for turnout in the 1999 general, 2000 general, and 2001 primary elections, whether the respondent is a Democrat, is a Republican, is an Independent, is black, is white, city of residence, and missing data on age; and integers for the respondent’s age and the number of family members in the household.
Table 2: 2SLS and ICSW estimates for Albertson and Lawrence (2009). Bootstrap 95% confidence intervals in parentheses. Dependent variables are knowledge (coded on a scale from 1-4 from least to most informed) and opinion (coded 1 if the respondent supported the proposition and 0 otherwise). Covariates include: television news-watching habits (coded on a seven point scale from never watches to watches everyday), newspaper reading habits (coded on seven point scale from never reads to reads everyday), interest in politics and national affairs (coded on a four point scale from low interest to high interest), party ID (coded on an 11 point scale), income (coded on a scale from 1 to 11 from poorest to richest), female (coded 1 if the respondent is female and 0 otherwise), education (coded on a 13 point scale from least to most educated) and white (coded 1 if the respondent is white and 0 otherwise).
Supporting Information Online Appendix A: Simulation Studies

In this online appendix, we perform exhaustive simulation studies to shed light on the following issues:

1. Optimal Choice of $\alpha$ for Winsorization
2. Properties of ICSW with and without Misspecification
3. Properties of Permutation Test with and without Misspecification
4. Comparisons of Multiple Test Statistics for Permutation Test Under Proper Specification
5. Bootstrap Confidence Interval Coverage

Baseline Data Generating Process

The following parameters define the “baseline” data generating process (DGP). Note the very flexible specification, which allows for a wide variety of compliance scores and relationships between covariates, compliance and heterogeneity.

\[ X_{1i}, X_{2i}, X_{3i} \sim U(-1, 1) \]
\[ a, b, c, d, e, f, g, h \sim U(-2, 2) \]
\[ AC_i = I[2 + a + bX_{1i} + cX_{2i} + dX_{3i} + N(0, 3)] \]
\[ A_i = I[-2 + e + fX_{1i} + gX_{2i} + hX_{3i} + N(0, 3)]AC_i \]
\[ C_i = AC_i - A_i \]
\[ D_i = D_iZ_i + A_i \]
\[ Y_{0i} = 5AC_i + N(0, 1) \]
\[ \tau_i = 5 \times (1 + X_{1i} + X_{2i} + X_{3i}) \]
\[ Y_i = Y_{0i} + \tau_i D_i \]

We observe and use \(X_{1i}, X_{2i}, X_{3i}\) as both covariates in 2SLS, and for the compliance score model.

**Misspecified Data Generating Processes**

**Heterogeneous Treatment Effects, Conditioning on Random Noise**

Same as baseline, except:
\[ X_{1i}', X_{2i}', X_{3i}' \sim U(-1, 1) \]
are the observed covariates, which are entirely independent from \(X_{1i}, X_{2i}, X_{3i}\).

**Homogeneous Treatment Effects, Conditioning on Random Noise**

Same as the previous DGP, except:
\[ \tau_i = 5, \forall i. \]

**Heterogeneous Treatment Effects, Model is Incorrect**

Same as the baseline DGP, except:
\[
AC_i = I[2 + a + (1 + bX_{1i})(1 + cX_{2i}) + dX_{3i}^3/3 + \text{Logistic}(0, 3)]
\]
\[
A_i = I[-2 + e + (1 + fX_{1i})(1 + gX_{2i}) + hX_{3i}^3/3 + \text{Logistic}(0, 3)]AC_i
\]
\[ \tau_i = 5 \times (1 + X_{1i} + X_{2i} + X_{1i}X_{2i} + X_{3i}) \]

**Heterogeneous Treatment Effects, Insufficient Conditioning Set**

Same as baseline, except only \(X_{1i}\) is observed.
(1) Optimal Choice of $\alpha$ for Winsorization

Assuming that the model is correct, we present plots of RMSE vs. $\alpha$ for $N \in 500, 1000, 2500, 5000, 10000, 25000$, generated using 2,500 iterations each, in Figures A1 and A2. We subclassify each of these plots to DGPs with low ($\text{mean}(P_C) < 0.33$), medium ($0.33 \leq \text{mean}(P_C) < 0.50$) and high ($0.50 \leq \text{mean}(P_C)$) compliance.
Figure A1: RMSE vs. $\alpha$ for $N \in 500, 1000, 2500$. 

A4
Figure A2: RMSE vs. $\alpha$ for $N \in 5000, 10000, 25000$. 

A5
Properties of ICSW with and without Misspecification

We assume $\alpha = 0.275$, as recommended in the text. We present tables of RMSE vs $N \in 500, 1000, 2500, 5000, 10000, 25000$ under the 5 different models, including misspecification. The permutation test is performed using SSR as the test statistic.

ICSW is compared to 3 other estimators: OLS (with $Z_i$), OLS (with $D_i$), and 2SLS. We again use 2,500 iterations for each $N$. We present these results in Figure A3. These results are also presented in tabular form in Tables A1-A5.
Figure A3: Properties of ICSW with and without Misspecification
<table>
<thead>
<tr>
<th>N</th>
<th>OLS (Z)</th>
<th>OLS (D)</th>
<th>2SLS</th>
<th>ICSW</th>
<th>Avg. Perm Test p</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2.54</td>
<td>2.53</td>
<td>0.92</td>
<td>0.72</td>
<td>0.01</td>
</tr>
<tr>
<td>1000</td>
<td>2.52</td>
<td>2.54</td>
<td>0.85</td>
<td>0.46</td>
<td>0.00</td>
</tr>
<tr>
<td>2500</td>
<td>2.51</td>
<td>2.52</td>
<td>0.77</td>
<td>0.27</td>
<td>0.00</td>
</tr>
<tr>
<td>5000</td>
<td>2.50</td>
<td>2.54</td>
<td>0.74</td>
<td>0.20</td>
<td>0.00</td>
</tr>
<tr>
<td>10000</td>
<td>2.50</td>
<td>2.53</td>
<td>0.72</td>
<td>0.14</td>
<td>0.00</td>
</tr>
<tr>
<td>25000</td>
<td>2.51</td>
<td>2.53</td>
<td>0.73</td>
<td>0.09</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table A1: Right

<table>
<thead>
<tr>
<th>N</th>
<th>OLS (Z)</th>
<th>OLS (D)</th>
<th>2SLS</th>
<th>ICSW</th>
<th>Avg. Perm Test p</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2.50</td>
<td>2.70</td>
<td>1.05</td>
<td>1.10</td>
<td>0.52</td>
</tr>
<tr>
<td>1000</td>
<td>2.50</td>
<td>2.67</td>
<td>0.89</td>
<td>0.92</td>
<td>0.50</td>
</tr>
<tr>
<td>2500</td>
<td>2.50</td>
<td>2.67</td>
<td>0.80</td>
<td>0.81</td>
<td>0.50</td>
</tr>
<tr>
<td>5000</td>
<td>2.50</td>
<td>2.69</td>
<td>0.75</td>
<td>0.76</td>
<td>0.49</td>
</tr>
<tr>
<td>10000</td>
<td>2.49</td>
<td>2.68</td>
<td>0.72</td>
<td>0.72</td>
<td>0.50</td>
</tr>
<tr>
<td>25000</td>
<td>2.50</td>
<td>2.72</td>
<td>0.74</td>
<td>0.74</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table A2: Noise (Heterogeneous)

<table>
<thead>
<tr>
<th>N</th>
<th>OLS (Z)</th>
<th>OLS (D)</th>
<th>2SLS</th>
<th>ICSW</th>
<th>Avg. Perm Test p</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2.50</td>
<td>2.64</td>
<td>0.53</td>
<td>0.58</td>
<td>0.52</td>
</tr>
<tr>
<td>1000</td>
<td>2.50</td>
<td>2.62</td>
<td>0.37</td>
<td>0.39</td>
<td>0.50</td>
</tr>
<tr>
<td>2500</td>
<td>2.49</td>
<td>2.61</td>
<td>0.23</td>
<td>0.23</td>
<td>0.50</td>
</tr>
<tr>
<td>5000</td>
<td>2.49</td>
<td>2.64</td>
<td>0.16</td>
<td>0.16</td>
<td>0.49</td>
</tr>
<tr>
<td>10000</td>
<td>2.47</td>
<td>2.61</td>
<td>0.11</td>
<td>0.11</td>
<td>0.50</td>
</tr>
<tr>
<td>25000</td>
<td>2.48</td>
<td>2.63</td>
<td>0.07</td>
<td>0.07</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table A3: Noise (Homogeneous)

A8
<table>
<thead>
<tr>
<th>(N)</th>
<th>OLS ((Z))</th>
<th>OLS ((D))</th>
<th>2SLS</th>
<th>ICSW</th>
<th>Avg. Perm Test (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2.97</td>
<td>2.85</td>
<td>0.94</td>
<td>0.83</td>
<td>0.15</td>
</tr>
<tr>
<td>1000</td>
<td>2.97</td>
<td>2.85</td>
<td>0.75</td>
<td>0.55</td>
<td>0.06</td>
</tr>
<tr>
<td>2500</td>
<td>2.97</td>
<td>2.87</td>
<td>0.60</td>
<td>0.36</td>
<td>0.01</td>
</tr>
<tr>
<td>5000</td>
<td>2.97</td>
<td>2.86</td>
<td>0.53</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>10000</td>
<td>2.95</td>
<td>2.85</td>
<td>0.51</td>
<td>0.22</td>
<td>0.00</td>
</tr>
<tr>
<td>25000</td>
<td>2.96</td>
<td>2.86</td>
<td>0.47</td>
<td>0.18</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table A4: Wrong Model

<table>
<thead>
<tr>
<th>(N)</th>
<th>OLS ((Z))</th>
<th>OLS ((D))</th>
<th>2SLS</th>
<th>ICSW</th>
<th>Avg. Perm Test (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2.55</td>
<td>2.65</td>
<td>0.99</td>
<td>0.92</td>
<td>0.11</td>
</tr>
<tr>
<td>1000</td>
<td>2.53</td>
<td>2.65</td>
<td>0.90</td>
<td>0.79</td>
<td>0.06</td>
</tr>
<tr>
<td>2500</td>
<td>2.51</td>
<td>2.62</td>
<td>0.79</td>
<td>0.69</td>
<td>0.03</td>
</tr>
<tr>
<td>5000</td>
<td>2.50</td>
<td>2.66</td>
<td>0.74</td>
<td>0.64</td>
<td>0.01</td>
</tr>
<tr>
<td>10000</td>
<td>2.50</td>
<td>2.65</td>
<td>0.73</td>
<td>0.62</td>
<td>0.00</td>
</tr>
<tr>
<td>25000</td>
<td>2.51</td>
<td>2.64</td>
<td>0.73</td>
<td>0.62</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table A5: Only X1
(3) Properties of Permutation Test with and without Misspecification

Both “conditioning on noise” models are equivalent for the permutation test. In Tables A1-A5, we report average $p$-values, by using one permutation for each of the 2,500 iterations. In Figure A4, we present the same SSR permutation test results graphically.

Figure A4: Properties of Permutation Test with and without Misspecification
(4) Comparisons of Multiple Test Statistics for Permutation Test Under Proper Specification

We assume correct specification. We use 200 iterations for the permutation test and 1000 iterations each for $N \in 150, 300, 450, 600$. We compare 4 different test statistics. The first test statistic, detailed in the text, is SSR. The second is SAR (sum of absolute residuals), or $\sum_{i=1}^{N} |D_i - \Pr(D_i = 1|X)|$. The third is the likelihood of the compliance score ML estimator. The fourth is simply the variance of the estimated $\Pr(D_i)$. In Figure A5, we show that the first three test statistics are all approximately equivalent for (correctly) rejecting the null, and that the variance performs considerably worse than the other three test statistics. In Table A6, we present the average $p$-values associated with each test statistic for each $N$. Lower $p$-values indicate better performance.

<table>
<thead>
<tr>
<th>$N$</th>
<th>SSR</th>
<th>SAR</th>
<th>Likelihood</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0.125</td>
<td>0.121</td>
<td>0.128</td>
<td>0.258</td>
</tr>
<tr>
<td>300</td>
<td>0.041</td>
<td>0.038</td>
<td>0.038</td>
<td>0.157</td>
</tr>
<tr>
<td>450</td>
<td>0.021</td>
<td>0.020</td>
<td>0.017</td>
<td>0.107</td>
</tr>
<tr>
<td>600</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.071</td>
</tr>
</tbody>
</table>

Table A6: Average $p$-values for Permutation Test
Figure A5: $p$-value Distributions
(5) Bootstrap Confidence Interval Coverage

In Table A7, we present % coverage with the bootstrap for $N \in 500, 2500, 5000$ assuming that the model is correct and $\alpha = 0.275$. Due to computational limitations, we use only 500 bootstrap iterations and 500 simulations for each $N$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>90% Interval Coverage</th>
<th>95% Interval Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.94</td>
<td>0.98</td>
</tr>
<tr>
<td>2500</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>5000</td>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table A7: Bootstrap Coverage
Supporting Information Online Appendix B: Permutation Test Plots

We present permutation test plots for both applications in Figures A6 and A7. For Green, Gerber and Nickerson (2003), all permutations are performed within randomization strata. For Albertson and Lawrence (2009), all permutations are performed without restrictions. Vertical line indicates observed SSR.

Figure A6: SSR Permutation Test for Green, Gerber and Nickerson (2003).
Albertson and Lawrence (2009) Permutation Test

Figure A7: SSR Permutation Test for Albertson and Lawrence (2009).
Supporting Information Online Appendix C: Compliance Score / Covariate Correlation Matrices

In Tables A8 and A9, we present compliance score and covariate correlation matrices.
<table>
<thead>
<tr>
<th></th>
<th>C. Score</th>
<th>Fam. Size</th>
<th>White</th>
<th>Black</th>
<th>Voted 00</th>
<th>Voted 99</th>
<th>Voted Pri.</th>
<th>Age</th>
<th>Age M.</th>
<th>Dem</th>
<th>Rep</th>
<th>Ind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp. Score</td>
<td>1.00</td>
<td>0.37</td>
<td>0.61</td>
<td>0.32</td>
<td>0.40</td>
<td>0.43</td>
<td>0.25</td>
<td>0.23</td>
<td>-0.55</td>
<td>0.38</td>
<td>0.31</td>
<td>-0.47</td>
</tr>
<tr>
<td>Columbus</td>
<td>-0.55</td>
<td>-0.17</td>
<td>-0.20</td>
<td>-0.07</td>
<td>-0.15</td>
<td>-0.05</td>
<td>-0.19</td>
<td>-0.00</td>
<td>1.00</td>
<td>-0.11</td>
<td>-0.03</td>
<td>0.79</td>
</tr>
<tr>
<td>Detroit</td>
<td>0.05</td>
<td>0.03</td>
<td>-0.30</td>
<td>-0.11</td>
<td>-0.00</td>
<td>-0.24</td>
<td>0.23</td>
<td>0.16</td>
<td>-0.23</td>
<td>-0.28</td>
<td>-0.19</td>
<td>-0.18</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>-0.45</td>
<td>-0.24</td>
<td>-0.21</td>
<td>-0.08</td>
<td>-0.06</td>
<td>-0.14</td>
<td>-0.01</td>
<td>-0.14</td>
<td>-0.16</td>
<td>-0.20</td>
<td>-0.14</td>
<td>-0.13</td>
</tr>
<tr>
<td>Raleigh</td>
<td>0.75</td>
<td>0.26</td>
<td>0.89</td>
<td>0.33</td>
<td>0.13</td>
<td>0.17</td>
<td>-0.00</td>
<td>0.02</td>
<td>-0.22</td>
<td>0.46</td>
<td>0.43</td>
<td>-0.18</td>
</tr>
<tr>
<td>St. Paul</td>
<td>0.07</td>
<td>0.15</td>
<td>-0.18</td>
<td>-0.07</td>
<td>0.16</td>
<td>0.30</td>
<td>0.06</td>
<td>-0.08</td>
<td>-0.14</td>
<td>-0.17</td>
<td>-0.12</td>
<td>-0.11</td>
</tr>
<tr>
<td>Fam. Size</td>
<td>0.37</td>
<td>1.00</td>
<td>0.23</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
<td>0.03</td>
<td>-0.05</td>
<td>-0.17</td>
<td>0.09</td>
<td>0.10</td>
<td>-0.14</td>
</tr>
<tr>
<td>White</td>
<td>0.61</td>
<td>0.23</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.13</td>
<td>0.19</td>
<td>-0.00</td>
<td>0.04</td>
<td>-0.20</td>
<td>0.33</td>
<td>0.47</td>
<td>-0.16</td>
</tr>
<tr>
<td>Black</td>
<td>0.32</td>
<td>0.09</td>
<td>-0.10</td>
<td>1.00</td>
<td>0.02</td>
<td>-0.00</td>
<td>0.00</td>
<td>-0.04</td>
<td>-0.07</td>
<td>0.32</td>
<td>0.04</td>
<td>-0.06</td>
</tr>
<tr>
<td>Voted 00</td>
<td>0.40</td>
<td>0.07</td>
<td>0.13</td>
<td>0.02</td>
<td>1.00</td>
<td>0.29</td>
<td>0.31</td>
<td>0.16</td>
<td>-0.15</td>
<td>0.08</td>
<td>0.09</td>
<td>-0.24</td>
</tr>
<tr>
<td>Voted 99</td>
<td>0.43</td>
<td>0.07</td>
<td>0.19</td>
<td>-0.00</td>
<td>0.29</td>
<td>1.00</td>
<td>0.16</td>
<td>0.16</td>
<td>-0.04</td>
<td>0.15</td>
<td>0.15</td>
<td>-0.09</td>
</tr>
<tr>
<td>Voted Pri.</td>
<td>0.25</td>
<td>0.03</td>
<td>-0.00</td>
<td>0.00</td>
<td>0.31</td>
<td>0.16</td>
<td>1.00</td>
<td>0.23</td>
<td>-0.19</td>
<td>-0.07</td>
<td>-0.01</td>
<td>-0.17</td>
</tr>
<tr>
<td>Age</td>
<td>0.23</td>
<td>-0.05</td>
<td>0.04</td>
<td>-0.04</td>
<td>0.16</td>
<td>0.16</td>
<td>0.23</td>
<td>1.00</td>
<td>-0.00</td>
<td>0.07</td>
<td>0.02</td>
<td>-0.00</td>
</tr>
<tr>
<td>Age Miss.</td>
<td>-0.55</td>
<td>-0.17</td>
<td>-0.20</td>
<td>-0.07</td>
<td>-0.15</td>
<td>-0.04</td>
<td>-0.19</td>
<td>-0.00</td>
<td>1.00</td>
<td>-0.11</td>
<td>-0.03</td>
<td>0.79</td>
</tr>
<tr>
<td>Dem</td>
<td>0.38</td>
<td>0.09</td>
<td>0.33</td>
<td>0.32</td>
<td>0.08</td>
<td>0.15</td>
<td>-0.07</td>
<td>0.07</td>
<td>-0.11</td>
<td>1.00</td>
<td>-0.15</td>
<td>-0.14</td>
</tr>
<tr>
<td>Rep</td>
<td>0.31</td>
<td>0.10</td>
<td>0.47</td>
<td>-0.04</td>
<td>0.09</td>
<td>0.15</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.15</td>
<td>1.00</td>
<td>-0.10</td>
</tr>
<tr>
<td>Ind</td>
<td>-0.47</td>
<td>-0.14</td>
<td>-0.16</td>
<td>-0.06</td>
<td>-0.24</td>
<td>-0.09</td>
<td>-0.17</td>
<td>-0.00</td>
<td>0.79</td>
<td>-0.14</td>
<td>-0.10</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table A8: Compliance Score / Covariate Correlation Matrix for Green, Gerber and Nickerson (2003). City dummies omitted from columns.
<table>
<thead>
<tr>
<th></th>
<th>Comp. Score</th>
<th>Party ID</th>
<th>Interest</th>
<th>Watch News</th>
<th>Educ.</th>
<th>Read News</th>
<th>Female</th>
<th>Income</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp. Score</td>
<td>1.00</td>
<td>0.37</td>
<td>0.65</td>
<td>0.28</td>
<td>0.36</td>
<td>0.62</td>
<td>-0.29</td>
<td>0.01</td>
<td>0.12</td>
</tr>
<tr>
<td>Party ID</td>
<td>0.37</td>
<td>1.00</td>
<td>-0.11</td>
<td>-0.02</td>
<td>-0.03</td>
<td>-0.05</td>
<td>0.04</td>
<td>-0.14</td>
<td>-0.22</td>
</tr>
<tr>
<td>Political Interest</td>
<td>0.65</td>
<td>-0.11</td>
<td>1.00</td>
<td>0.17</td>
<td>0.25</td>
<td>0.36</td>
<td>-0.06</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>Watch News</td>
<td>0.28</td>
<td>-0.02</td>
<td>0.17</td>
<td>1.00</td>
<td>0.03</td>
<td>0.15</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Education</td>
<td>0.36</td>
<td>-0.03</td>
<td>0.25</td>
<td>0.03</td>
<td>1.00</td>
<td>0.26</td>
<td>-0.10</td>
<td>0.30</td>
<td>0.15</td>
</tr>
<tr>
<td>Read News</td>
<td>0.62</td>
<td>-0.05</td>
<td>0.36</td>
<td>0.15</td>
<td>0.26</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>Female</td>
<td>-0.29</td>
<td>0.04</td>
<td>-0.06</td>
<td>-0.07</td>
<td>-0.10</td>
<td>-0.10</td>
<td>1.00</td>
<td>-0.11</td>
<td>0.00</td>
</tr>
<tr>
<td>Income</td>
<td>0.01</td>
<td>-0.14</td>
<td>0.16</td>
<td>-0.03</td>
<td>0.30</td>
<td>0.23</td>
<td>-0.11</td>
<td>1.00</td>
<td>0.08</td>
</tr>
<tr>
<td>White</td>
<td>0.12</td>
<td>-0.22</td>
<td>0.20</td>
<td>0.02</td>
<td>0.15</td>
<td>0.17</td>
<td>0.00</td>
<td>0.08</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table A9: Compliance Score / Covariate Correlation Matrix for Albertson and Lawrence (2009).