A Bayesian Approach to Inference with Instrumental Variables: Improving Estimation of Treatment Effects with Weak Instruments and Small Samples

Adriana Crespo-Tenorio
Department of Political Science
Washington University in St. Louis

Jacob M. Montgomery
Department of Political Science
Washington University in St. Louis

July 10, 2013

ABSTRACT

Standard methods for estimating the causal effects of endogenous regressors rely on assumptions not always appropriate for political science applications. In settings with small samples or weak instruments, textbook approaches such as two-stage least-squares and limited information maximum likelihood can produce inefficient and even inconsistent estimates of causal effects. In this paper, we implement a fully Bayesian approach to estimating causal effects in an instrumental variables framework as outlined by Chib (2003). We show that this algorithm more reliably recovers treatment effects of interest than standard approaches in the context of small samples and weak instruments. Moreover, this framework is easily adapted to accommodate binary outcomes. We assess the performance of the technique using synthetic data and apply it to the study of the effect of political institutions on government spending and the effect of education on political participation.

*We wish to thank Kosuke Imai, Matthew Gabel, Don Green, and Brendan Nyhan for helpful comments. We are grateful to Rachel Milstein Sondheimer and Donald P. Green for generously sharing their data. A previous version of this paper was presented at the 2013 meeting of the Midwest Political Science Association in Chicago, IL.
Identifying and understanding causal relationships is one of the central tasks of social science research. However, in many areas, hypotheses cannot be tested with randomized treatments and laboratory-levels of control. Scholars must instead draw inferences based on either observational data or experiments conducted with inherently imperfect levels of control and compliance. As a consequence, a great deal of attention has gone into developing methods to generate accurate and precise estimates of causal effects in the presence of unmeasured confounders.

Confounders – variables that correlate with both the outcome and the explanatory variable – are of particular interest since, when overlooked, they increase Type II error rates and bias estimates of causal effects. Causal inference assumes that the treatment is randomly assigned to individuals such that the only factor that is systematically different between treatment and control groups is the treatment itself. Confounding clearly violates this assumption and prevents the identification of the true effect. One approach adopted by scholars attempting to deal with this violation is to estimate an instrumental variables (IV) model which can identify causal effects in the presence of confounding, with the most common methods being two-stage least-squares (2SLS) or some form of limited information maximum likelihood (LIML).

While these techniques have become increasingly common in political science research (c.f., Sovey and Green 2010, 194), important limitations remain. In particular, the assumptions underlying standard approaches are not appropriate in situations frequently encountered in real-world empirical applications. First, inference with these models relies on the asymptotic properties of the estimators which need not apply in finite samples. For instance, 2SLS estimates have been shown to be biased in the context of moderately weak instruments and finite samples (Angrist and Pischke 2009; Bound et al. 1995). Second, when the instrument in question is not strongly related to the endogenous regressor of interest – in other words, the instrument is “weak”– the asymptotic theory behind methods such as 2SLS breaks down completely, leading to not only biased but inconsistent and inefficient estimates (Betz 2013; Staiger and Stock 1997).
Finally, these standard approaches have proven to be somewhat inflexible and difficult to adjust to handle many common data-types encountered in applied research. Of particular interest are instances where both the endogenous regressor and the outcome variable are dichotomous. In these situations, due to the nonlinearity in the data generating process, 2SLS and LIML estimators are guaranteed only to provide estimate the local average treatment effects (LATE) even when other assumptions of large sample sizes and strong instruments are met. Other quantities of interest, especially average treatment effect (ATE) and the effect of the treatment on the treated (TOT) require additional distributional assumptions. To handle these issues some authors have suggested a bivariate probit model (e.g., Angrist and Pischke 2009) in conjunction with a more agnostic estimation technique such as 2SLS. However, Freedman and Sekhon (2010) show that the maximum-likelihood bivariate probit algorithms in common software suites are numerically unstable and recommend against their use.

In this paper, we implement a Bayesian model for estimating treatment effects with unmeasured confounders developed by Chib (2003), which addresses many of these concerns (see also Greenberg 2013). It provides more accurate estimates of treatment effects in the context of small samples and weak instruments. Moreover, it is possible to extend the model to handle binary outcomes, and it is straightforward to calculate a full posterior for treatment effects of interest.

The paper proceeds as follows. We begin by describing the most common approaches to instrumental variable analysis, noting that Bayesian models are relatively scarce in this literature. In Section 3, we present a Bayesian IV model for both continuous and binary outcomes and identify some important advantages to adopting this approach. Using simulated datasets, we then show how the Bayesian technique yields significantly less biased and more efficient estimates than traditional approaches under a variety of conditions including weak instruments and small samples. Finally, we apply this model to the Persson and Tabellini (2004) dataset that explores the effect of electoral institutions on government spending and the Sondheimer and Green (2010) dataset that estimates the effect educational attainment on voter turnout. We conclude with a discussion of possible extensions to the basic model.
2. INSTRUMENTAL VARIABLE ANALYSIS WITH AN ENDOGENOUS DICHOTOMOUS REGRESSOR

We consider the problem of identifying a causal effect in the context of the potential outcomes framework introduced by Rubin (1974) and carefully linked to the assumptions of instrumental variables by Angrist et al. (1996). Let $y$ be a continuous outcome, $T \in \{0, 1\}$ be a binary treatment variable of theoretical interest, and $x$ be a set of covariates (including a constant). In the potential outcomes framework, for each subject $i \in [1, 2, \ldots, n]$, we observe an outcome $y_i$ which is a realization of one of two potential outcomes $(y_{i0}, y_{i1})$ corresponding to the two possible levels of treatment, $(T_{i0}, T_{i1})$. Thus, $y_{i0}$ and $y_{i1}$ are random variables with means $E(y_{0}) = \beta + x_i \gamma_2$, $E(y_{1}) = x_i \gamma_2$, and

$$
y_{i0} = x_i \gamma_2 + \eta_i,
$$

$$
y_{i1} = \beta + x_i \gamma_2 + \eta_i,
$$

(1)

where $\eta_i$ is an unobserved random variable with mean 0. The observed outcome, $y_i$, is given by $y_i = (1 - T_i)y_{i0} + T_i y_{i1}$ and the average treatment effect (ATE) is $E(y_{1}) - E(y_{0}) = \beta$, or the difference between the average outcomes given different levels of treatment.

Assuming that the unobserved errors are conditionally independent of the treatment, $\eta_i \perp T_i|x_i$, the ATE can be estimated using standard regression techniques (Rosenbaum and Rubin 1983). However, in many settings, this assumption does not hold; some factors may affect both the treatment assignment and the outcome. In this scenario, OLS estimates are known to be inconsistent.

2.1. Traditional methods of estimation in IV analysis

To generate consistent estimates of the true causal effect of $T$, the instrumental variables approach introduces a set of variables, $z$, that determine $T$ but are independent of $y$ conditioned on the covariates $x$. With some additional assumptions,\(^1\) it is then possible to apply methods from struc-

\(^1\)The additional required assumptions are the stable unit treatment value assumption (SUTVA), monotonicity, and the exclusion restriction (Angrist et al. 1996).
tural equation modeling to generate consistent and efficient estimates of causal effects. The basic approach is to set up and solve a system of simultaneous equations,

\[ T_i = x_i' \gamma_1 + z_i' \pi + \varepsilon_{1i} \] (2)
\[ y_i = x_i' \gamma_2 + T_i \beta + \varepsilon_{2i} \] (3)

where \((\varepsilon_{1i}, \varepsilon_{2i})' \sim N(0, \Omega)\). The level of confounding is measured by the correlation coefficient, \(\rho = \frac{\omega_{12}}{\sqrt{\omega_{11} \omega_{22}}}\). (Note that the binary nature of \(T_i\) is ignored in this model, a point we return to below.)

Two-stage least-squares estimation and its limitations: Far and away, the dominant approach to estimating this model is two-stage least-squares (2SLS), which is a method attractive for both its simplicity and robustness (Angrist and Pischke 2009). We first estimate \(\hat{T}_i = x_i' \hat{\gamma}_1 + z_i' \hat{\pi}\), where \(\hat{\gamma}_1\) and \(\hat{\pi}\) are OLS estimates from Equation (2). Taking \(V_i = [x_i', \hat{T}_i]'\), the vector of coefficients for the second stage is then

\[
[\hat{\gamma}_2', \hat{\beta}] = \left[ \sum_{i=1}^{n} V_i V_i' \right]^{-1} \sum_{i=1}^{n} V_i y_i.
\] (4)

Under standard regression assumptions, this estimator is distributed normally around \([\hat{\gamma}_2, \hat{\beta}]\) with covariance matrix estimated consistently by \(\sum_{i=1}^{n} V_i V_i' \left[ \sum_{i=1}^{n} V_i V_i' \right]^{-1} \hat{\omega}_{22}\).

As comprehensively reviewed in standard texts (c.f., Angrist and Pischke 2009; Wooldridge 2002), 2SLS is an intuitive and robust way of estimating causal effects in a wide variety of situations. Nonetheless, this now-standard IV estimation technique has at least two important limitations (Kleibergen and Zivot 2003). First, 2SLS estimators are asymptotically biased with weak instruments and have a nonstandard (possibly bimodal) distribution (Bound et al. 1995; Staiger and Stock 1997). Second, the estimation itself depends on asymptotic properties which may not be appropriate for small-\(n\) studies.\(^2\) In combination, these limitations mean that 2SLS can lead to wildly incorrect estimates and overly large confidence intervals in the presence of weak instruments and

\(^2\)Of less practical importance, 2SLS can provide more precise but more biased estimated of \(\beta\) when additional weak instruments are introduced. Moreover, in multi-stage settings, 3SLS is not invariant to the ordering (Kleibergen and Zivot 2003).
finite samples.

Second, 2SLS can sometimes prove to be quite limiting and inflexible. Of particular interest are IV analyses with dichotomous outcomes such as voting turnout, vote-choice, or voter registration. Here, due to the nonlinearity of the response variable, even under ideal circumstances 2SLS provides only an estimate of the local average treatment effect (LATE), or the effect of the treatment on compliers. Other quantities of interest, such as the average treatment effect (ATE) and the effect of the treatment on the treated (TOT) are unavailable.\(^3\)

**Limited information maximum likelihood:** A number of alternative estimation techniques have been forwarded in the literature to address these problems. Most prominently, the limited information maximum likelihood (LIML) estimator is now the standard recommendation for applied analysts when dealing with weak instruments (e.g., Flores-Lagunes 2007; Gawande and Li 2009; Sovey and Green 2010; Staiger and Stock 1997). The LIML estimator is a \(\kappa\)-class estimator based on a single equation (Greene 2000; Staiger and Stock 1997). More formally, LIML estimator is

\[
[\hat{\gamma}'_2, \hat{\beta}] = \left[ V(I - \kappa M_z) V' \right]^{-1} \left[ V(I - k M_z) y \right], \tag{5}
\]

where \(k\) is the smallest eigenvalue of \(WP_z W'[W(M_z)W']^{-1} W_i = [T_i, y_i]'\), \(M_z = I - P_z\) and \(P_z\) is the projection of \(z\). The variance-covariance matrix is identical to 2SLS (Greene 2000, 687).

However, closer examination of the properties of LIML show it to be no panacea. To begin with, if the model in Equations (2) and (3) is exactly identified, it can be shown that the LIML estimator (Equation (5)) reduces to the 2SLS estimator (Equation (4)). In other words, with only one instrument (the usual case in political science applications), LIML estimates will be identical to 2SLS (Greene 2000, 681–687). Further, in the presence of weak instruments, LIML estimates of \(\beta\) are also inconsistent (their finite sample properties with non-weak instruments are that it they are median unbiased). Indeed, it has previously been shown that the first and second moments of \(\kappa\)-class estimators like LIML are undefined in finite samples (Sawa 1972).

\(^3\)To be clear, these quantities of interest are unavailable **even if the researcher is willing to make additional distributional assumptions.**
Fuller (1977) proposed a modification to $\kappa$-class estimators to ensure the existence of the first two moments, with the same limiting distribution as the original estimators.\(^4\) Previous simulations (c.f. Flores-Lagunes 2007) have shown that Fuller’s adjustment to LIML performs somewhat better with weak instruments, and it is this estimate we refer to as “LIML” in our results below.\(^5\)

**Bivariate probit:** One plausible solution is to estimating an IV model with a binary outcome and a binary treatment is to estimate a bivariate probit model. In this model, using $1(\cdot)$ as the standard indicator function, we replace Equations (2) and (3) with

\[
T_i = 1(x_i'\gamma_1 + z_i\pi + \varepsilon_{1i} > 0) \\
y_i = 1(x_i'\gamma_2 + T_i\beta + \varepsilon_{2i} > 0)
\]

where $\varepsilon_{1i}$ and $\varepsilon_{2i}$ are distributed as a bivariate normal distribution with ones along the diagonal. Given accurate parameter estimates from this model, the bivariate normal assumption allows us to make estimates of the ATE and TOT.

On its face, this seems a reasonable strategy. Angrist and Pischke, for instance, state that, “Bivariate probit probably qualifies as harmless in the sense that it’s not very complicated and easy to get right using packaged software routines” (Angrist and Pischke 2009, p 201). However, they further recommend that bivariate probit should be used primarily in conjunction with a 2SLS estimation of the LATE. Moreover, Freedman and Sekhon (2010) show through Monte Carlo simulations that standard maximum likelihood routines in statistical package such as *Stata* produce estimates that are far from the true MLE. They generously conclude that estimates from these package are “questionable.” Thus, the standard advice to applied researchers is to simply ignore the dichotomous nature of both the treatment and the outcome and to use 2SLS.

\(^4\)We replace $\kappa$ in Equation (5) with $\kappa - \frac{\alpha}{n^p}$, where $n$ is the sample size, $P$ is the number of instruments, $\alpha$ is an arbitrary constant (Flores-Lagunes 2007; Fuller 1977). Fuller (1977) suggested setting $\alpha = 1$ to minimize bias and $\alpha = 4$ to minimize the estimator’s mean square error (Fuller 1977, 951). In our examples below, we set $\alpha = 1$.

\(^5\)These are, of course, not the only alternatives in this extensive literature. Imbens and Rosenbaum (2005) provide a permutation approach for inference with weak instruments that has been extended by Betz (2013). Chamberlain and Imbens (2004) provides a random effects approach for estimating IV models in the presence of a large number of instruments. Flores-Lagunes (2007) provide a more comprehensive review of several additional alternatives and compares them via simulation.
2.2. Bayesian approaches

The issues discussed above strongly suggest that a Bayesian approach could be of great benefit. First, prior structures frequently make efficient estimation of weakly identified models quite easy. Since proper prior distributions guarantee proper posteriors, the likelihood function and its identification are less important for deriving estimates in Bayesian models (Imbens and Rubin 1997). Second, Bayesian methods have successfully been applied to generate unbiased and consistent estimates in many situations where standard maximum-likelihood methods are unsatisfactory or poorly behaved (c.f. Clinton et al. 2004; Gill 2008).

However, Bayesian research on instrumental variable models is quite limited and recent relative to the immense literature from the classical perspective. This is somewhat surprising as Angrist et al. (1996) applied the Rubin causal model to IV analysis in overtly Bayesian terms. Kleibergen and Zivot (2003) trace Bayesian IV analysis back to Dreze (1976), which developed a Bayesian framework for LIML estimation, although causal inference was not at the center of much of this early work. Imbens and Rubin (1997) proposed a phenomenological Bayesian model to estimate causal effects in experiments with non-compliance (see also Frangakis et al. 2002; Hirano et al. 2000). However, this method’s approach to estimation is fundamentally different from simply estimating parameters for the model in Equations (2) and (3) as we do here.

The approach we examine below has several advantages relative to these alternatives. First, most of the models above assume both a continuous treatment and a continuous outcome, while the Chib (2003) model applies data augmentation (Albert and Chib 1993) to address the binary

---

6 In addition to those paper cited elsewhere, other notable contributions to the Bayesian IV literature includes Chib and Hamilton (2000), Chib and Hamilton (2002), and Chib (2007). Heckman et al. (2013) implements a Bayesian model using a latent-trait approach. Other presentations of Bayesian IV analysis include Gelman and Hill (2006, p 510) and Rossi et al. (2005).

7 Thus, they argue that it is possible to think of the potential outcomes as two unknown values that follow a joint probability distribution (Angrist et al. 1996, 446).

8 Kleibergen and Zivot (2003) show that Bayesian versions of both 2SLS and LIML are less sensitive to the introduction of auxiliary weak instruments.

9 To begin, it is based on dividing observations into strata based on compliance profiles (Frangakis et al. 2002). We set aside a more detailed comparison of the Chib (2003) and Imbens and Rubin (1997) methods. Our purpose here is more narrow in seeking to focus on alternative measures of estimating the simple structural model above rather than comparing all possible ways of estimating treatment effects with non-compliance.
nature of treatments. Further, by explicitly accounting for the binary nature of the treatment, the Chib model is more straightforward to conceptualize within the potential outcomes framework facilitating the direct calculations of quantities such as the LATE. Finally, in this framework it is straightforward to handle binary treatments with binary outcomes (Chib and Greenberg 1998).

3. BAYESIAN IV ESTIMATION WITH ENDOGENOUS BINARY TREATMENTS

In this section, we present the Chib (2003) model in terms of a continuous outcome and binary treatment. We then briefly discuss the extension of the model to binary outcomes and the calculation of treatment effects with heterogeneous treatment effects. In the following sections, we then evaluate the model using both synthetic data and replicate two prominent studies that implement IV methods.

3.1. Model Specification

We begin by making two assumptions common to IV estimation:

**Assumption 1** \( \text{COV}(z, T | x) > 0 \)

**Assumption 2** \( z \perp \epsilon_2 | x, T \)

Assumption 1 states that the instrument has a conditional effect on the endogenous treatment. Assumption 2 states that knowledge about the data-generating process of the instrument will not contribute to our knowledge about the outcome except through the treatment.

We then assume an outcome model and treatment assignment mechanism,

\[
    T_i = 1(x_i' \gamma_1 + z_i \pi + \epsilon_{1i} > 0), \quad (7)
\]

\[
    y_i = x_i' \gamma_2 + T_i \beta + \epsilon_{2i}, \quad (8)
\]
with \( f(\varepsilon_{1i}, \varepsilon_{2i}) = N(0, \Omega) \). For identification, we assume that \( \omega_{12} = \omega_{21} \) and \( \omega_{11} = 1 \). Note that here we are explicitly accounting for the binary nature of the treatment. Applying this normality assumption to the potential outcomes in Equation (1), we obtain the marginal densities

\[
\begin{align*}
\text{f}_0(y_i | x_i, \gamma_1, \gamma_2, \beta, \Omega) &= N(x_i' \gamma_2, \omega_{22}) \\
\text{f}_1(y_i | x_i, \gamma_1, \gamma_2, \beta, \Omega) &= N(x_i' \gamma_2 + \beta, \omega_{22})
\end{align*}
\]  

(9)

Letting \( \Psi = (\gamma_1, \gamma_2, \beta, \pi, \Omega) \), from Equation (7) we have the marginal distribution of \( T \),

\[
\text{Pr}(T_i = 1 | z_i, x_i, \Psi) = \Phi(x_i' \gamma_1 + z_i \pi) \text{ where } \Phi \text{ is the CDF of the standard normal distribution.} \]

The likelihood can then be expressed as the product of the two marginal distributions

\[
L(y, T | x, z, \Psi) = \prod_{i \in F} f_0(y_i | x_i, \Psi) \text{Pr}(T_i = 0 | x_i, z_i, \Psi) \times \prod_{i \in G} f_1(y_i | x_i, \Psi) \text{Pr}(T_i = 1 | x_i, z_i, \Psi),
\]

(10)

where \( F \) is the set where we observe \( T_i = 0 \), and \( G \) represents the subset where \( T_i = 1 \).

To complete the model, it is necessary to specify priors on the parameters. Anticipating the estimation technique below, we redefine the parameters in the model such that \( \theta_1 = \omega_{22} - \omega_{12}^2 \) and \( \theta_2 = \omega_{12} \) with \( \theta = [\theta_1, \theta_2] \). Constraining \( \theta_1 \) to be positive ensures that \( \Omega \) will be positive definite. For convenience, we also define \( B = [\gamma_1', \pi, \gamma_2', \beta] \). Finally, we use vague conjugate priors \( \pi(B) = N(B_0, \Sigma_B_0) \), \( \pi(\theta_2 | \theta_1) = N(m_0, \theta_1 M_0) \), and \( \pi(\theta_1) = IG(v_0/2, d_0/2) \), where \( IG(\cdot) \) is the inverse gamma distribution with mean \( (v_0/2)((d_0/2) - 1)^{-1} \).

3.2. Estimation

Since the treatment is binary, the likelihood function is not tractable. We therefore use data augmentation, as proposed by Albert and Chib (1993), which to reformulates the model in terms of a latent, continuous variable for treatment, \( T^* \). The assumption is that \( T_i = 1(T^*_i > 0) \), where

\[10\]In a maximum-likelihood framework, it seems intuitive to implement some form of 2SLS using a probit or logit regression in the first stage. However, this is the so-called "forbidden regression," and will lead to incorrect inferences in a frequentist or maximum likelihood framework (Angrist and Pischke 2009; Wooldridge 2002).
\( T^* \sim N(x_i'\gamma_1 + z_i\pi, 1). \)

Using the latent treatment variable, the joint model (for outcome and treatment assignment) now has two continuous variables that can model jointly with a bivariate normal density. For convenience, we define \( y^* = (y, T^*). \) In addition, let

\[
X_i = \begin{pmatrix} x_i' & T_i & 0' & 0 \\ 0' & 0 & x_i' & z_i \end{pmatrix},
\]

which allows us to write the conditional distribution of the augmented data as

\[
f(y_i^* | X, B, \Omega) \propto \exp[((y_i^* - X_iB)'\Omega^{-1}(y_i^* - X_iB)]. \tag{11}\]

The model is now a bivariate linear regression and the full conditional distributions for both the augmented data \( T^* \) and the regression parameters \( B \) can be explicitly derived. The full MCMC algorithm then has only three steps. First, we sample the augmented data \( T^* \) from the truncated normal distribution

\[
T_i^* | X, B, \Omega \sim \begin{cases} 
N(x_i'\gamma_1 + z_i\pi + (\omega_{12}/\omega_{22})(y_i - x_i'\gamma_2), \theta_1/\omega_{22}); I(-\infty, 0) & \text{if } T_i = 0 \\
N(x_i'\gamma_1 + z_i\pi + (\omega_{12}/\omega_{22})(y_i - x_i'\gamma_1 - \beta), \theta_1/\omega_{22}); I(0, \infty) & \text{if } T_i = 1 
\end{cases} \tag{12}\]

Second, we sample the vector of regression coefficients \( B \) from

\[
N(\hat{B}, \hat{\Sigma}_B), \tag{13}\]

where \( \hat{\Sigma}_B = (\Sigma_{B_0}^{-1} + \sum X_i'\Omega^{-1}X_i)^{-1} \) and \( \hat{B} = \hat{\Sigma}_B(\Sigma_{B_0}^{-1}B_0 + \sum X_i'\Omega^{-1}y_i^*). \) In step three, we sample \( \theta \) using a Metropolis-Hastings step with a tailored non-central t-distribution as the proposal density (Chib and Greenberg 1998). This is discussed in Appendix A.
3.3. Discussion: Contrasts with traditional estimation methods

To preview the results below, the Bayesian IV model as proposed by Chib (2003) generally performs better – sometimes dramatically better – than either 2SLS or LIML estimators with weak instruments and finite samples. As we show, the Bayesian model is often substantially more accurate and efficient, and is rarely outperformed by the traditional approaches. So it is worth pausing here to consider how the Chib model contrasts with the standard approach. Specifically, we discuss why the estimates can sometimes be so different despite the models appearing (superficially) quite similar and the interpretive advantages of adopting the Bayesian framework.

Why are the Bayesian estimates so different?: There are at least three important differences in how the Bayesian IV model operates relative to the traditional methods discussed above. First, unlike the 2SLS or LIML, the MCMC algorithm is iterative. This means that the uncertainty inherent in the estimation of the first-stage parameters propagates through the model. Thus, in the presence of weak instruments, the significant uncertainty about the estimate of $\pi$ and $\omega_{12}$ is correctly accounted for in the second stage regression on $y$. This contrasts strongly with, for instance, 2SLS where a single point estimate for $\hat{\pi}$ is calculated, and our uncertainty about its correct value is ignored. Moreover, the iterative nature means that information flows in two directions as the outcome $y^* = (y, T^*)$ is modeled jointly. Thus, the Bayesian estimation process goes beyond borrowing information about the origins of the treatment to correct estimates in the second stage; it also uses information in the second stage (e.g., $y_i$) to improve estimates of our model of $T_i$.

Second, the Bayesian model directly incorporates the binary nature of the treatment variable. While this comes at the cost of some additional assumptions, it also has the advantage of preventing the model from providing nonsensical predictions outside of the $[0, 1]$ interval. It also allows for heterogeneity in the way that the instrument affects the instrument via the nonlinear link function.

Third, and perhaps most fundamentally, the Bayesian approach for inference about the parameter of interest (in this case $\beta$) is to update prior beliefs through the likelihood in order to estimate
Thus, we can calculate a proper posterior even if identification is weak. As Jackman (2009, p 440-441) notes, poor identification “poses no formal problem in a Bayesian analysis. Identification (or the absence of it) is a property of the likelihood function, whereas Bayesian inference simply uses the likelihood function to map through the data from prior beliefs to posterior beliefs.” This stands in stark contrast with traditional estimation techniques where the unbiasedness, efficiency, and consistency of the estimator are fundamentally based upon the identification of the model (i.e., a strong instrument) and asymptotic distributional properties. When neither of these conditions hold, the estimators and associated confidence intervals can be wildly off.

**Further advantages:** Before turning to our simulations, there are two additional advantages to the Bayesian approach that are worth noting. To begin with, having a full posterior for all parameters improves interpretability. So, for instance, we are able to directly estimate a posterior for the level of confounding in the model as \( \rho = \frac{\omega_1^2}{\omega_2^2} \). Thus, we can calculate a credible interval for the level of confounding in the model. We can also directly estimate causal effects of interest. The most obvious is the average treatment effect, which is defined as

\[
ATE = \mathbb{E}(y_1) - \mathbb{E}(y_0) = \mathbb{E}(f_1(y|X,B,\Omega)) - \mathbb{E}(f_0(y|X,B,\Omega)) = \beta. 
\] (14)

We can then directly estimate the posterior density of the ATE as the posterior density of \( \beta \). Since the assumed joint distribution of \( y^* \) implies constant effects, this is also our estimate for the TOT and LATE. However, it is also possible to calculate these quantities when the assumed joint distribution is non-linear. As an example, the calculations for generating a posterior for these causal effects in the bivariate probit model are discussed in Appendix B.

A final point of contrast between the Bayesian model and its traditional alternatives, is that it is relatively straightforward to extend this model to accommodate different data configurations. For instance, we can model binary outcomes in this framework as follows. The treatment assignment

---

11 We emphasize here that our point here is in discussion the prior structure, not the priors themselves. Even with weak identification, the model will provide almost identical results for a wide array of reasonably diffuse priors. This point is illustrated using the political participation example in Appendix C.

12 Of course, if identification is non-existent interpretation of the posterior will be very difficult.
model remains unchanged, but we re-define the outcome model in Equation (8) as

\[ y_i^* = x_i' \gamma_2 + T_i \beta + \epsilon_{2i} \]
\[ y_i = 1(y_i^* > 0), \]  

so that \( y^* = (y^*, T^*) \) follows a bivariate normal density. In the MCMC algorithm, this adjustment simply requires that we sample the augmented outcome, \( y^* \), in addition to the existing steps. Appendix A provides additional details.\(^{13}\)

The discussion above illustrates how the Chib model differs from standard methods for estimating IV models in terms of estimation, interpretability, and flexibility. However, the question remains as to whether it will aid researchers in practice to better analyze data and understand causal relationships. In the next sections, therefore, we turn to comparing the Bayesian and traditional IV estimation techniques using synthetic data and two real-world examples.

**4. SIMULATION RESULTS**

In this section, we apply the Bayesian IV framework to synthetic data. Our goal is to demonstrate two major points. First, we provide some exemplar evidence that the Bayesian model accurately and precisely recovers treatment effects of interest even in the context of small samples and weak instruments, and that these estimates are more accurate and more efficient than either LIML or standard 2SLS estimates. Second, we conduct a more thorough comparison of the performance of the Chib (2003) model to traditional IV methods using a broader set of simulated datasets. Focusing more narrowly on unbiased estimation, this simulation study again show that the Bayesian model significantly outperform traditional approaches.

\(^{13}\)We discuss further extensions of the basic model in the concluding section.
Throughout, the simulations assume the true data generating process to be

\[
T^*_i = x_i \gamma_1 + z_i \pi + \varepsilon_{i1} \\
y_i = x_i \gamma_2 + T_i \beta + \varepsilon_{2i},
\]

(16)

with parameter values set as shown in Table 1. This corresponds exactly with Equations (7-8). Our measure of instrument strength is analogous to the estimate in Chib (2003) and Flores-Lagunes (2007), and is termed below as a pseudo R-squared estimate. Using the augmented data for the treatment \((T^*)\) as the dependent variable, we measure instrument strength as the percent of the total variation explained by the instruments. Following the results provided by Hahn and Hausman (2002), this suggests that:

\[
R^2_z = \frac{\pi' E[z'z] \pi}{\pi' E[z'z] \pi + 1}.
\]

Using \(p \in [1, 2, \ldots, P]\) to indicate the instrument,\(^{14}\) and dividing the explanatory power equally among all instruments, this implies that \(\pi_p = \sqrt{R^2_z/(P(1 - R^2_z))}\). Our measure of confounding is \(\rho = \omega_{12}/\sqrt{\omega_{22}}\), which is the correlation between the outcome variable \(y\) and the augmented treatment data \(T^*\). Since we assume in our simulations that \(\omega_{22} = 1\), this implies that \(\rho = \omega_{12}\). Setting \(\omega_{22} = 1\) also facilitates interpretation as the estimate of the treatment effect, \(\beta\), indicates the number of standard deviations change associated with receiving treatment. In all simulations, we set \(\beta = 1\), suggesting a strong effect.

### 4.1. Exemplar simulations

We begin by focusing on a limited set of simulations to illustrate the advantages of the Bayesian approach relative to traditional methods. Figure 1 compares these estimation techniques for different levels of instrument strength, \(R^2_z = (0.01, 0.05, 0.1, 0.25, 0.5, 0.75)\), with a moderately small sample size, \(n = 100\), and significant level of confounding, \(\rho = 0.7\).\(^{15}\) For each level of instrument

\(^{14}\)In over-identified models, we may have more than one instrument. In the more common setting with one instrument and one endogenous regressor, this implies that \(P = 1\).

\(^{15}\)In addition, we assume normal errors and a single instrument. See Table 1 for further information.
strength, we fit the 2SLS, LIML, and Bayesian models and calculate the point estimate and 95% confidence or credible interval. We also calculate the point estimate and 95% confidence interval for a single-stage ordinary least squares (OLS) model to serve as a reference.

These panels in Figure 1 shows the point estimates and 95% CI for each estimator. There are two aspects of the figure that we wish to emphasize. First, the bias in the estimation of the ATE is much lower for the Bayesian method. For instance, the Bayesian estimate is almost entirely unbiased when $R^2_z = 0.1$, while the 2SLS and LIML estimates are off by over three. Given that the ATE is measured in terms of standard deviations of the dependent variable, this represents both a significant and a substantial difference. Moreover, the bias with very weak instruments ($R^2_z < 0.25$) for 2SLS and LIML are sometimes worse (in terms of absolute bias) than the OLS estimates.

Second, the interval estimates for the Bayesian methods are much more reasonable, especially with weaker instruments. That is, the Bayesian methods provide much more information about the true causal effect than the conventional estimators. Indeed, the interval estimates for 2SLS and LIML frequently go beyond a plausible three standard deviations in each direction. This reflects an inability of these methods to learn adequately from the data to pin down the treatment effect when identification is even moderately weak.

Figure 2 explores how the different estimators work as a function of level of confounding. We simulate using a sample size of $n = 100$ and a fairly weak instrument, $R^2_z = 0.1$. We vary only the level of confounding in the data generating process, $\rho = (0.25, 0.5, 0.7, 0.9)$. Again, each panel shows the point estimates and 95% CI for the ATE estimates. For this simulated data, the 2SLS and LIML estimators are again quite poorly behaved. We can again see that they both are outperformed by the Bayesian alternatives in terms of bias and efficiency. In particular, note again that traditional estimation techniques have unreasonably large confidence intervals, ranging to as much as 14 standard deviations.\(^\text{16}\) Not also that for all levels of confounding, the confidence intervals for the 2SLS and LIML confidence intervals include zero.

\(^{16}\)The coverage of the Bayesian model is actually a bit small for high levels of confounding (e.g., $\rho = .9$). This will contribute to higher Type-II errors if the estimates are used for hypothesis testing. However, this is actually quite reasonable behavior. With a weak instrument, severe confounding, and a small sample, there simply is not sufficient information for correctly updating posterior estimates of the treatment effect.
Figure 1: IV estimators and confidence intervals by instrument strength (n=100, ρ=0.7)

The panels show the point estimates and 95% CI for the estimators of treatment effects for ordinary least squares (OLS), two-stage least-squares, LIML, and Bayesian estimation method. Note that the Bayesian estimators are less biased and the interval estimates are better calibrated for weak instruments.
These panels show the point estimates and 95% CI for the estimators of treatment effects for ordinary least squares (OLS), two-stage least-squares, LIML, and Bayesian estimation method. Note that the Bayesian estimators are less biased and the interval estimates are better calibrated for nearly all levels of confounding.
Table 1: Parameters for simulation study of Bayesian and traditional IV estimators. The model for the simulated data is shown in Equation (16).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of confounding</td>
<td>$\rho = \omega_{12}$</td>
<td>0.25, 0.5, 0.7, 0.9</td>
</tr>
<tr>
<td>Instrument strength (Pseudo R-Squared)</td>
<td>$R_z^2$</td>
<td>0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9</td>
</tr>
<tr>
<td>Number of instruments</td>
<td>$P$</td>
<td>1, 3</td>
</tr>
<tr>
<td>Sample size</td>
<td>$n$</td>
<td>100, 500, 1000</td>
</tr>
<tr>
<td>First-stage control coefficients</td>
<td>$\gamma_1$</td>
<td>(-1, 1.5)</td>
</tr>
<tr>
<td>Second-stage control coefficients</td>
<td>$\gamma_2$</td>
<td>(1, -0.75)</td>
</tr>
<tr>
<td>Treatment coefficient</td>
<td>$\beta$</td>
<td>1</td>
</tr>
<tr>
<td>Controls</td>
<td>$x_i$</td>
<td>$(1, N(0, 1))$</td>
</tr>
<tr>
<td>Instrument(s)</td>
<td>$z_i$</td>
<td>$z_i \sim N(0, I)$</td>
</tr>
<tr>
<td>Variance terms</td>
<td>$\omega_{11}, \omega_{22}$</td>
<td>1, 1</td>
</tr>
<tr>
<td>Error distribution</td>
<td></td>
<td>$N(0, \Omega), Mvt-t(scale=\Omega, df = 12)$</td>
</tr>
<tr>
<td>Replications</td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

4.2. Systematic simulations

We now turn to a more systematic examination of the comparative performance of the Bayesian and traditional estimators. We again simulate data according to Equation (16), but now consider a more complete set of parameter values, modeled on the comprehensive simulations provided in Flores-Lagunes (2007). Table 1 shows the parameter values considered and the method for generating the synthetic data. We vary the level of confounding, the instrument strength, the number of instruments and the sample size. To consider the sensitivity of the model to distributional assumptions, we also vary the distribution of the error terms $(\varepsilon_1, \varepsilon_2)$. We draw these errors from either a multivariate normal distribution, $N(\mathbf{0}, \Omega)$, or a multivariate t-distribution, $Mvt-t(scale=\Omega, df = 12)$. For each parameter setting, we generated six datasets, for a total of over 2,200 analyses.\footnote{In future versions of this paper, we hope to modestly increase the number of simulations at each setting to ten.}

For each dataset, we calculate the bias for the estimated treatment effect in using the Bayesian, LIMLS, and 2SLS estimators. Figure 3 provides the results. The top panels plot the bias for each estimator against instrument strength. The second panel shows how each estimator performs as a
function of the level of confounding. The bottom panel explores how estimator performance varies as a function of sample size.

There are two main points illustrated by Figure 3. First, the Bayesian methods outperform the traditional 2SLS and LIML estimators in nearly every setting. Second, the performance of the 2SLS and LIML estimators (or lack thereof) appears to depend almost entirely on the strength of the instrument. The bottom two panels show that these estimators can be wildly wrong even with low levels of confounding and large sample sizes in the presence of weak instruments. This is somewhat surprising, as it means that even large sample sizes and modest levels of confounding are not sufficient for estimating causal effects in the absence of quite strong instruments. In contrast, the Bayesian methods become more accurate as sample size increases and as confounding decreases as we would normally expect.

5. EMPIRICAL APPLICATIONS

We now turn to applying the Bayesian IV model to two prominent studies, each providing an excellent opportunity for demonstrating advantages of the Bayesian approach as they involve small sample sizes, weak instruments, or both. The first is the Persson and Tabellini (2004) analysis of the effect of specific democratic institutions on the size of government and welfare spending. This example illustrates the capacity for Bayesian IV analysis to improve the precision for estimates of causal effects, even with modest ($n < 100$) samples. Second, we re-analyze the Sondheimer and Green (2010) study of the effect of education on political participation. This example illustrates the flexibility of the Bayesian approach for handling non-normal outcomes, and shows how the approach offers more reasonable estimates in the context of weak instruments.
Figure 3: Comparison of IV estimators by instrument strength, confounding, and sample size

Each panel shows the bias for the Bayesian, 2SLS, and LIML estimator of the treatment effect. The Bayesian estimates are generally more accurate estimators of the causal effect.
5.1. The Effect of Institutions on the Size of Government Size and Redistribution

We begin by applying the Bayesian IV approach to study the effects of electoral rules and form of government on the size of government and redistribution levels. We use a cross-national sample of 80 democratic countries gathered and analyzed in Persson and Tabellini (2004) (henceforward PT). Since the dataset features a small sample size and a moderately weak instruments, the simulations above suggest the Bayesian approach should produce a more accurate analysis.

Regime type and electoral rule are widely considered key determinants of economic policy because they affect the distribution of political responsibility in a country (Lewis-Beck and Stegmaier 2000; Powell 2000) and thus parties’ incentives to use economic policy for electoral gain (Franzese 2002; Lizzeri and Persico 2001). Studies have consistently found that proportional representation (PR) systems are associated with higher redistribution levels relative to majoritarian systems (e.g., Iversen and Soskice 2006; Persson et al. 2007), which we term the electoral rules effect. Likewise, parliamentary systems have has been shown to increase the size of government and lead to higher levels of redistribution relative to presidential systems (e.g., Tsebelis 1995). Persson and Tabellini (2005) attribute this empirically-supported assertion to the absence of checks and balances in parliamentary regimes as well as incentives to make compromises in the legislature in order to hold together government coalitions. We term this the form of government effect.

However, establishing the direction of causality in these relationships is empirically challenging since there are many possible confounding factors. Several other observable and unobservable factors may simultaneously explain differences in institutional layouts and fiscal policy outcomes across countries. For example, as PT point out, colonial history could explain a country’s institutional characteristics. (Most of the presidential countries were also former Spanish colonies.) At the same time, colonial history may play a role in determining the redistribution structures.

PT explore the question using a cross-national design. In this widely cited paper, PT subject their data to a number of analyses, not all of which we reproduce here, including several 2SLS regressions. In general, their results show support for the idea that majoritarian systems and pres-
idential systems lead to smaller governments and that PR systems lead to higher levels of redistribution. However, PT find insufficient evidence to support the claim that parliamentary systems lead to higher levels of redistribution.

PT measure the size of government as government expenditures as a percentage of Gross Domestic Product (GDP). Similarly, they use welfare spending as a percentage of GDP as a measure for the composition of government spending. The treatments are two dichotomous indicators, equal to unity if the regime is presidential and if the electoral system is majoritarian. Thus, the expected treatment effects of both institutions on both dependent variables is negative.

These endogenous institutions are instrumented by indicators of the epoch when the current constitutional rules were established, under the assumption that constitutional history can explain cross-country variation in constitutional rules, but are unrelated to current levels of spending.\textsuperscript{18} Table 2 shows how the instruments are related to these endogenous institutions. The table shows that presidential regimes occur more frequently in constitutions that were established most recently, suggesting a relatively strong instrument. Indeed, the R-squared estimates for the independent effect of the instruments is over 0.2. However, these instruments are somewhat less strong in predicting the electoral rules. Specifically, the difference in the adjusted R-squared estimates when the instruments are included is only 0.04 for the government size sample and 0.08 for the welfare spending sample. Furthermore, with only 80 countries and six possible institution/constitution establishment combinations, within-category sample sizes are very small.

We now turn to comparing the estimates of the treatment effects estimated from 2SLS and LIML models compare with those produced by the Bayesian model described above.\textsuperscript{19} Our goal is to estimate the effect of each endogenous institution (form of government and electoral rules) on government size and welfare spending separately, using constitution establishment times as

\textsuperscript{18}For our purposes, we set aside issues concerning the validity of these instruments to focus on how the Bayesian IV estimation differs from traditional 2SLS.

\textsuperscript{19}We run three independent chains of the MCMC algorithm for 100,000 iterations. Our prior distribution for the coefficients is centered at zero with a variance of five. For $\omega_{12}$, our prior is a truncated normal distribution with mean zero, and its variance equals 0.5, and the prior for $\omega_{22}$ is an Inverse Gamma distribution with shape $= 1/4$ and rate $= 1$.\textsuperscript{20} The first half of the posterior was discarded as burn-in, and convergence was assessed for the remaining half using the superdiag library in R (Tsai and Gill 2012).
Table 2: Description of institutions and constitutional origin from Persson and Tabellini (2004)

<table>
<thead>
<tr>
<th>Constitution Origin</th>
<th>Electoral Rule</th>
<th>Form of Government</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PR</td>
<td>Majoritarian</td>
</tr>
<tr>
<td>pre–1921</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>9.30%</td>
<td>5.81%</td>
</tr>
<tr>
<td>1921–1950</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>9.30%</td>
<td>1.16%</td>
</tr>
<tr>
<td>1951–1980</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>13.95%</td>
<td>13.95%</td>
</tr>
<tr>
<td>post–1980</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>30.23%</td>
<td>16.28%</td>
</tr>
<tr>
<td>Total (N=86)</td>
<td>54</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>62.79%</td>
<td>37.21%</td>
</tr>
</tbody>
</table>

Percentages may not add to 100. Persson and Tabellini (2004) explain that for Costa Rica, Sri Lanka, Cyprus and France, “the electoral rule and the form of government originate in different periods, and for these countries the indicator variables for both periods take a value of 1.” (Persson and Tabellini 2004, 28)

instruments. We then re-estimate these models using the Bayesian IV framework. Note that we standardize the outcome variables to speed posterior convergence.

The results, shown in Table 3, provide several important insights that support the conclusions main theoretical argument in Persson and Tabellini (2004) and illustrate the advantages of utilizing the Bayesian model described above. The Bayesian model performs better than the 2SLS and LIML models in the sense that our estimates are more precise, despite the small sample. In all cases, the ATE estimates differ only marginally from 2SLS, and are generally somewhat lower. However, the credible intervals for the Bayesian analyses are uniformly smaller than the comparable confidence intervals. For instance, the estimate for the 2SLS estimate of the effect of majoritarianism on the size of government is $-1.080$ with a standard error of 0.846, which does not reach traditional levels of statistical significance. The Bayesian posterior mean, however, is $-0.913$ with a posterior standard deviation of 0.356, indicating that the 95% credible interval will not include zero.

Thus, there are also important substantive implications of these results. As in Persson and

---

21These results do not represent an exact replication of Persson and Tabellini (2004), who include both endogenous regressors in the same estimation equation and also include a large number of other covariates. For illustrative purposes, we simplify their analysis considerably.
Table 3: 2SLS and Bayesian IV results for the effect of form of government (presidential = 1, parliamentary = 0) and electoral rules (majoritarian = 1, proportional representation=0) on government size and welfare spending

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Size of Government</th>
<th>Welfare Spending</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Form of Government</td>
<td>Electoral Rule</td>
</tr>
<tr>
<td></td>
<td>2SLS</td>
<td>Fuller's LIML</td>
</tr>
<tr>
<td>Constitution Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.200)</td>
<td>(0.200)</td>
</tr>
<tr>
<td>1921–1950</td>
<td>0.105</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>(0.185)</td>
<td>(0.185)</td>
</tr>
<tr>
<td>1951–1980</td>
<td>0.319</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>(0.160)</td>
<td>(0.160)</td>
</tr>
<tr>
<td>post-1980</td>
<td>0.586</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>(0.185)</td>
<td>(0.185)</td>
</tr>
<tr>
<td>Government Size Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.432</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>(0.332)</td>
<td>(0.339)</td>
</tr>
<tr>
<td>ATE</td>
<td>-1.116</td>
<td>-1.120</td>
</tr>
<tr>
<td></td>
<td>(0.536)</td>
<td>(0.551)</td>
</tr>
<tr>
<td>N</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>ρ</td>
<td>0.141</td>
<td>0.128</td>
</tr>
</tbody>
</table>

Outcome variables have been standardized. The coefficients are 2SLS and Fuller’s LIML estimates and Bayesian posterior means. In parenthesis, we report standard errors for the 2SLS and Fuller’s LIML columns and posterior standard deviations for the Bayesian columns. Variables included but not reported in the model include the age of the country’s democracy and an indicator for British colonial past. At the Constitution Level, 2SLS and LIML estimates are at a linear scale, while Bayesian estimates are on a probit scale. ρ, the level of confounding, is calculated using the posterior means of ω_{22} and ω_{12}. For Fuller’s LIML models, we set α = 1.
Tabellini (2004), the 2SLS estimates for the effect of electoral rules on government size and welfare spending are statistically indistinguishable from zero. In contrast, the Bayesian estimates strongly support the hypothesis that parliamentary systems spend more than majoritarian systems. In addition, the degree of estimated confounding is small. The point estimates for $\rho$ are all modest, ranging from 0.128 to 0.150.

5.2. The Effect of Education on Voter Turnout

We now apply the Bayesian IV approach to an important question in American politics: Does educational attainment cause higher turnout? We use the dataset from Sondheimer and Green (2010), which combines three experiments that increase educational attainment with turnout data in federal elections. This example is apt for demonstrating the power of the Bayesian approach as the dataset exhibits three features common to political science applications, but which are poorly handled by conventional methods: small sample sizes, weak instruments, and binary outcomes. Moreover, we are able to show how it is possible to pool inferences across multiple studies in the Bayesian framework.

That high socio-economic status, especially education, is strongly related to political participation is a well-established fact in American politics research (e.g., Rosenstone and Hansen 1993; Verba and Nie 1972; Verba et al. 1995). Yet the unique causal effect of education is difficult to establish despite its power in predicting participation (Schlozman 2002, 442). On the one hand, it is argued that through education, citizens acquire skills and resources that are necessary for political participation, and become more politically aware (Jackson 1995). On the other hand, education levels are highly correlated with other characteristics of citizens’ background which are also well-known correlates of both participation and educational attainment (Berinsky and Lenz 2011; Brady et al. 1995; Card 1999). Put differently, citizens’ motivation to vote may come from family or community background factors that also encourages them to attain a high-school education. Socioeconomic status is thus a confounding factor in the relationship between turnout and education.
making the unique contribution of educational attainment difficult to identify (Tenn 2007).

Sondheimer and Green (2010) (henceforward SG) recently addressed this problem using a unique dataset analyzed in an IV framework. SG leverage three randomized experiments aimed at stimulating better education in subjects that have been widely used in the social sciences (e.g., Barnett 2010; Hanushek 1999; Heckman et al. 2010; Kahne and Bailey 1999): the High/Scope Perry Preschool Experiment (Perry), the I Have A Dream program (IHAD), and the Student-Teacher Achievement Ratio (STAR) experiment. The basic approach is to consider these experiments as encouragement studies, where educational attainment is the actual causal variable of interest and the experimental assignments are instruments. SG then match data from participants to turnout in federal elections between 2000 and 2004, where the outcome variable is simply an indicator for voting in any election in this period.

On its face, this study appears to meet the two assumptions stated in Section 3. To the extent that the programs were effective, education is correlated with the instrument (the original experimental treatment). To the extent that the treatments were randomly assigned and administered before the participants were able to vote, treatment is exogenous to turnout. This makes the experimental treatments a theoretically sound instrument for education. SG conclude that each study provides mild evidence in favor of a causal relationship in isolation, while a weighted average of the estimated coefficient across studies suggests an unambiguous strong causal effect.

However, there are two features of this data, which is displayed in Table 4, that are particularly relevant for our purposes. First, the sample sizes are quite small for the Perry (N=123) and IHAD (N=58) experiments. With the four potential combinations of treatment and outcome, we are left with very few observations of each potential outcome. With 58 subjects in the IHAD sample, for example, we have only four subjects in the treatment group who did not complete high school.

Additionally, high school graduation rates in the STAR experiment were high both in absolute terms and in comparison to the other studies. Graduation rates were 90.08% in the treatment group and 84.97% in the control group. Intuitively, this suggests that the effect of being assigned given the experimental treatment status, the potential outcomes are (Diploma, Vote), (Diploma, No Vote), (No Diploma, Vote), (No Diploma, No Vote)
Table 4: Description of educational experiment data from Sondheimer and Green (2010)

<table>
<thead>
<tr>
<th></th>
<th>Perry</th>
<th>IHAD</th>
<th>STAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS Graduation</td>
<td>65.00%</td>
<td>78.95%</td>
<td>90.08%</td>
</tr>
<tr>
<td>Turnout</td>
<td>18.33%</td>
<td>42.11%</td>
<td>46.83%</td>
</tr>
<tr>
<td>N</td>
<td>60</td>
<td>19</td>
<td>252</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS Graduation</td>
<td>44.44%</td>
<td>61.54%</td>
<td>84.97%</td>
</tr>
<tr>
<td>Turnout</td>
<td>12.70%</td>
<td>33.33%</td>
<td>42.22%</td>
</tr>
<tr>
<td>N</td>
<td>63</td>
<td>39</td>
<td>559</td>
</tr>
</tbody>
</table>

to the treatment in the STAR program was quite small. In other words, there is low covariance between the instrument and the endogenous variable. For these reasons, traditional approaches to instrumental variables may be biased and even inconsistent.

We replicated the bivariate probit regression results from the analysis in SG. These estimates are shown in the columns labeled ML in Table 5. We then re-analyze this data using the Bayesian IV framework, and the results are shown in the columns labeled Bayesian.\(^{23}\) Broadly speaking, the Bayesian results in Table 5 confirm the conclusions from the maximum likelihood models, but there are important differences. First, the coefficients estimated in the Bayesian probit model differ considerably from the maximum-likelihood estimates. As we anticipated, in the STAR model – where we suspected a weaker instrument – the discrepancies between the two approaches are very obvious. At the voter turnout level, the Bayesian posterior mean for the coefficient measuring the impact of education on turnout is half the size and has larger standard deviations relative to the ML estimates. This pattern, where the Bayesian estimates for the coefficient are substantially smaller, also occurs in the IHAD and Perry models, where the sample sizes are small.

The Bayesian estimates for the effect of education on turnout are positive in all three experiments, which supports the idea that education levels increase the likelihood of voter turnout. How-

---

\(^{23}\)For each experiment, we run three independent chains of the MCMC algorithm for 100,000 iterations. Our prior distribution for the coefficients is centered at zero with a variance of ten. For \(\omega_{12}\), the level of confounding, our prior is a truncated normal distribution with mean zero, and its variance equals two. We discard 50% of the draws as burn-in and assess convergence using the `superdiag` library in R (Tsai and Gill 2012). As robustness checks, we re-estimated the models with vaguer prior distributions. The results are displayed in Appendix C.
Table 5: Bivariate probit regression and Bayesian IV results for the downstream effects of educational attainment on turnout

<table>
<thead>
<tr>
<th></th>
<th>Perry ML</th>
<th>Perry Bayesian</th>
<th>IHAD ML</th>
<th>IHAD Bayesian</th>
<th>STAR ML</th>
<th>STAR Bayesian</th>
<th>Pooled ML</th>
<th>Pooled Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS Graduation Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−0.140</td>
<td>−0.130</td>
<td>0.295</td>
<td>0.303</td>
<td>1.037</td>
<td>1.033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>0.529</td>
<td>0.505</td>
<td>0.513</td>
<td>0.497</td>
<td>0.246</td>
<td>0.258</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voter Turnout Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−1.602</td>
<td>−1.368</td>
<td>−1.013</td>
<td>−0.425</td>
<td>−1.850</td>
<td>−1.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS Graduation</td>
<td>1.084</td>
<td>0.660</td>
<td>1.051</td>
<td>0.132</td>
<td>1.948</td>
<td>0.970</td>
<td>1.401</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>123</td>
<td>123</td>
<td>58</td>
<td>58</td>
<td>811</td>
<td>811</td>
<td>992</td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>−0.095</td>
<td>−0.104</td>
<td>−0.218</td>
<td>−0.196</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATE</td>
<td>0.148</td>
<td>0.037</td>
<td>0.296</td>
<td>0.263</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOT</td>
<td>0.110</td>
<td>0.025</td>
<td>0.288</td>
<td>0.250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ML columns replicate Table 5 in Sondheimer and Green (2010). Coefficients are ML estimates and Bayesian posterior means. In parenthesis, we report bootstrapped standard errors for ML estimates and posterior standard deviations for Bayesian estimates.

However, the Bayesian posterior means suggest a much more modest effect with significantly more uncertainty. Beyond the coefficients, this is also reflected in the posterior estimates for the ATE and TOT, which are shown in the bottom rows of Table 5. In particular, the IHAD study shows a very modest effect for education, indicating that graduating high school is associated only with a 3-4% increase in turnout. (As we show below, the posteriors for the ATE and TOT are non-symmetric, so the standard deviation should be interpreted with caution.)

Likewise, the Bayesian estimates for the pooled causal effect across all three experiments are modest and there is significant uncertainty. SG construct a precision-weighted estimate of the coefficient measuring the effect of education on turnout on the probit scale. This is shown on the right column of Table 5 and suggests that, when pooling inferences across experiments, there is strong evidence in favor of the effect of high-school graduation on turnout. However, the Bayesian analysis only weakly supports this conclusion. Due to the non-linear nature of this
These posterior densities of the treatment effects for each dataset were obtained by calculating the ATE and TOT at each iteration of the MCMC algorithm. Based on the pooled posterior density for each, $P(ATE \leq 0 \approx 0.0775)$ and $P(TOT \leq 0 \approx 0.0834)$. For a full description of the derivation of these estimators in the bivariate probit framework, see Appendix B. The pooled estimates are an average of the estimated ATE and TOT at each MCMC draw weighted by sample size.

model, the posteriors for the treatment effects can be extremely non-symmetric. Figure 4 displays the posterior estimates for the ATEs (left panel) and TOT (right panel) for each study, as well as the pooled posterior density.\textsuperscript{24}

The posterior mean pooled ATE is 0.263, which suggests that the causal effect of high school graduation on voting in a federal election is a 26% increase – quite a significant effect. However, as suggested by the coefficients discussed above, there is significant uncertainty. The posterior probability that the pooled ATE is less than zero, $Pr(ATE \leq 0)$, is approximately 0.0775. Given the inherent limitations of the data (i.e., small sample sizes and weak instruments) we interpret this to provide modest support for the hypothesized causal relationship. However, these results are much weaker than those provided by maximum likelihood estimation. The right panel shows the posterior estimates for the effect of the treatment on the treated. The mean of the posterior suggests a treatment effect of an approximately 25% increase in turnout. In words, this means that if those who completed high school had failed to do so, the probability that they would have voted in a federal election would be 25% lower. This effect is sizable, but it there is once again substantial

\textsuperscript{24}Pooled estimates are calculated by taking making a sample-size weighted average from each draw from the posterior.
Table 6: 2SLS and LIML estimates for the downstream effects of educational attainment on turnout

<table>
<thead>
<tr>
<th>HS Graduation Level</th>
<th>Perry 2SLS</th>
<th>Fuller LIML</th>
<th>IHAD 2SLS</th>
<th>Fuller LIML</th>
<th>STAR 2SLS</th>
<th>Fuller LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.444</td>
<td>0.444</td>
<td>0.615</td>
<td>0.615</td>
<td>0.850</td>
<td>0.850</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.062)</td>
<td>(0.075)</td>
<td>(0.075)</td>
<td>(0.014)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.206</td>
<td>0.206</td>
<td>0.174</td>
<td>0.174</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>(0.089)</td>
<td>(0.088)</td>
<td>(0.132)</td>
<td>(0.132)</td>
<td>(0.026)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>Voter Turnout Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.005</td>
<td>0.019</td>
<td>0.023</td>
<td>0.153</td>
<td>-0.345</td>
<td>-0.225</td>
</tr>
<tr>
<td></td>
<td>(0.178)</td>
<td>(0.162)</td>
<td>(0.585)</td>
<td>(0.439)</td>
<td>(0.696)</td>
<td>(0.600)</td>
</tr>
<tr>
<td>HS Graduation</td>
<td>0.274</td>
<td>0.249</td>
<td>0.311</td>
<td>0.504</td>
<td>0.902</td>
<td>0.764</td>
</tr>
<tr>
<td></td>
<td>(0.320)</td>
<td>(0.292)</td>
<td>(0.879)</td>
<td>(0.646)</td>
<td>(0.804)</td>
<td>(0.693)</td>
</tr>
<tr>
<td>N</td>
<td>123</td>
<td>123</td>
<td>58</td>
<td>58</td>
<td>811</td>
<td>811</td>
</tr>
</tbody>
</table>

Coefficients are 2SLS/Fuller-corrected LIML estimates with standard errors an parentheses. Note that the estimates of the effect of graduation on turnout are implausibly large for the STAR and IHAD experiments. In the Fuller-corrected LIML models, we set $\alpha = 1$.

uncertainty: the probability that the TOT is less than 0 is roughly 0.0834.

For the sake of completeness, in Table 6 we display the 2SLS and LIML estimates using the Stata `ivreg2` command. In these cases, the LATE is simply the coefficient for high school graduation in the turnout level. Note that many of these estimates are prima facia implausible. For instance, in the STAR experiment, 2SLS estimates the effect of education as leading to 90% increase in the predicted probability of voting in a federal election. Note also, that the implied confidence interval fall well outside of the possible $[-1, 1]$ interval for both the STAR and IHAD experiments. This provides additional support for the notion that simply ignoring the binary nature of the outcome and treatment can lead to serious errors in estimation.

6. CONCLUSION

In this article, we implemented a Bayesian model for treatment effects with confounders as outlined by Chib (2003). We show that this method provides more accurate estimates of treatment effects than traditional methods under a number of conditions common to political science research. Specifically, the model outperforms 2SLS and LIML in the context of weak instruments
and small sample sizes. Moreover, the model is more easily adapted to handle a binary outcome and binary endogenous regressor simultaneously.

After discussing how weak instruments and small samples represent potential problems for standard instrumental variables analysis, we outlined the model and estimation strategy developed by Chib (2003). In addition to being more accurate and more flexible model, we showed that this MCMC approach allows for calculation of the full posterior for multiple treatment effects of interest with heterogeneous treatment effects as well as the level of confounding.

Using simulated data, we showed that this approach is helpful when sample sizes are small and instruments are weak – even when the data is not distributed as assumed. Finally, we applied this approach to an IV estimation of the effect of constitutional design on government spending (Persson and Tabellini 2004) and the effect of educational attainment and voter turnout (Sondheimer and Green 2010). These analyses illustrated how the Bayesian approach can provide more reliable and accurate estimates of treatment effects, and are more flexible for handling binary outcomes than traditional approaches.

Before concluding, it is worth noting several limitations and discuss fruitful directions for future extensions. The primary concern with the model as presented above is the restrictive prior structure imposed on the data. The model assumes a bivariate normal distribution for both the augmented treatment and the outcome. Moreover, the presentation above assumes that all covariates are related to outcomes in a simple additive manner.

However, one of the advantages of adopting the Bayesian framework is that it is possible to extend the model in a variety of ways, leveraging the extensive array of tools available in the Bayesian literature, to address these concerns. Chib and Greenberg (2007) provide a variant of the model used here under non-parametric assumptions, and Chib et al. (2009) extend this further to allow for non-random sample selection. However, although far more flexible, the models just cited also make specific distributional assumptions about the joint distribution of the treatment and the outcome that may not always be appropriate. In theory, however, the modeling framework can accommodate flexible joint distributions, such as Dirichlet process mixtures (Chib and Greenberg...
2010), although we are aware of no such implementations in the literature.

It should also be possible adapt this model to other data structures that are common in social science research. For example, the IV model for civil war and economic growth in Miguel et al. (2004) involves a binary outcome with continuous treatment, and the model in Mauro (1998) includes a continuous outcome with ordered categorical treatment. In addition, time series cross sectional studies are pervasive in political science research, and future variants may better account for panel structure and autocorrelation.

However, while we believe there is a wide scope for future extensions, the results above demonstrate that even this straightforward Bayesian implementation offers significant practical advantages over the standard techniques now dominant in the literature. With large samples and strong instruments, all approaches will provide nearly identical answers. Yet, researchers are seldom blessed with this kind of data. Under more realistic circumstances with finite samples and weak instruments, the Bayesian implementation we describe promises to aid researchers at making more accurate, efficient, and interpretable estimates of treatment effects in the context of unmeasured confounders.

**REFERENCES**


APPENDIX A: MCMC ESTIMATION

Let $y$ be a continuous outcome, $T$ be a binary treatment, $x$ be a matrix of covariates (including the constant), and $z$ be a binary instrument for $T$. The model assumes that

\[
T^*_i = x_i'\gamma_1 + z_i\pi + \epsilon_{1i} \\
T_i = 1(T^*_i > 0) \\
y_i = x_i'\gamma_2 + T_i\beta + \epsilon_{2i}
\]

where $(\epsilon_{1i}, \epsilon_{2i}) \sim N(0, \Omega)$, $\omega_{11} = 1$, and $\omega_{12} = \omega_{21}$. For convenience, we define $y^* = (T^*, y)$. In addition, let

$$X_i = \begin{pmatrix} x_i' & z_i & 0' & 0 \\ 0' & 0 & x_i' & T_i \end{pmatrix}$$

which allows us to write the conditional distribution of the augmented data as shown in Equation (11). Anticipating the estimation technique below, we redefine the parameters in the model such that $\theta_1 = \omega_{22} - \omega_{12}^2$ and $\theta_2 = \omega_{12}$ with $\theta = [\theta_1, \theta_2]$. The priors for the parameters are listed in the main text.

The MCMC algorithm then has three steps. First, we sample the augmented data $T^*$ from the truncated normal distribution as shown in Equation (12). Second, we sample the vector of regression coefficients $\beta$ as shown in Equation (13).

We sample $\theta$ using a Metropolis-Hastings step with a tailored non-central t-distribution as the proposal density (Chib and Greenberg 1998). The target density is

$$\pi(\theta | y^*, X, B, \Omega) \propto IG(\theta_1)(v_0/2, d_0/2)N(\theta_2)(m_0, \theta_1 M_0) \times \prod_{i=1}^n N(y_i)(x_i'\gamma_2 + T_i\beta + \theta_2(T_i^* - x_i'\gamma_1 - z_i\pi), \theta_1)$$

First, we sample a new value of $\theta$ (we denote the proposal value $\theta'$, and the current value as $\theta$) from the proposal density $q(\theta | y_i, T^*_i, z, \beta)$, the proposal density. Specifically, we use a bivariate T
with non-centrality parameter approximated by the mode of

\[
\prod_{i=1}^{n} N(y_i | x_i' \gamma_2 + T_i \beta + \theta_2 (T_i^* - x_i' \gamma_1 - z_i \pi), \theta_1).
\]  

(17)

The dispersion parameter is set equal to the inverse of the observed information matrix of Equation (17) when evaluated at the estimated mode. The degrees of freedom can be set to any arbitrary value, although this choice should affect only the speed of the sampling algorithm and not the estimates (we use \(df = 5\) in the applications in the main text).

We can calculate the modal values of \(\theta\) using \texttt{optim} function in \texttt{R}. Specifically, we maximize:

\[
-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\theta_1) - \sum_{i=1}^{n} (y_i - x_i' \gamma_2 + T_i \beta + \theta_2 (T_i^* - x_i' \gamma_1 - z_i \pi))^2
\]

with respect to \(\theta_1\) and \(\theta_2\). We keep the sampled value with probability

\[
\alpha = \min \left\{ \frac{\pi(\theta' | y_i, T_i^*, z, \beta) \cdot q(\theta | \theta', y_i, T_i^*, z, \beta)}{\pi(\theta | y_i, T_i^*, z, \beta) \cdot q(\theta' | \theta, y_i, T_i^*, z, \beta)} \right\}
\]

APPENDIX B: BAYESIAN IV ANALYSIS WITH BINARY OUTCOMES

\textit{MCMC estimation}

The continuous-outcome model can by modified to accommodate a binary outcome by stating that

\[
y_i = I(x_i' \gamma_2 + T_i \beta + \epsilon_2) \quad \text{and} \quad y^* = (T^*, y^*).
\]

Thus,

\[
\begin{pmatrix}
T^* \\
y^*
\end{pmatrix} = \begin{pmatrix} x_i & z_i & 0' & 0' \\
0' & 0 & x_i' & T_i
\end{pmatrix} \begin{pmatrix}
\gamma_2 \\
\beta \\
\gamma_1 \\
\pi
\end{pmatrix} + \begin{pmatrix} \epsilon_2 \\\n\epsilon_1 \end{pmatrix},
\]

where the observed outcomes arise as \(y_i = 1(y_i^* > 0)\) and \(T_i = 1(T_i^* > 0)\). The variance covari-
The MCMC algorithm has three basic steps: (1) drawing the augmented values of \( y^* \), (2) drawing values of the regression parameters \( \mathbf{B} \) and (3) drawing the value of \( \theta_2 \) using a Metropolis-Hastings step. First, we sample the augmented data \( y^*_i \) and \( T^*_i \). The latter is drawn from Equation (12) but substituting in \( y^*_i \) for \( y_i \). The former is drawn from

\[
\begin{aligned}
y^*_i | T^*_i, \mathbf{X}, \mathbf{B}, \Omega &\sim \begin{cases} 
N(\mathbf{x}'_i \gamma_2 + T_i \beta + \theta_2 (T^*_i - \mathbf{x}'_i \gamma_1 - z_i \pi), \theta_1); I(-\infty, 0) & \text{if } y_i = 0 \\
N(\mathbf{x}'_i \gamma_2 + T_i \beta + \theta_2 (T^*_i - \mathbf{x}'_i \gamma_1 - z_i \pi), \theta_1); I(0, \infty) & \text{if } y_i = 1
\end{cases}
\end{aligned}
\]

Next, we sample the regression parameters using Equation (13).

Finally, we sample \( \theta_2 \) using Metropolis-Hastings step. The target density is

\[
\pi(\theta_2 | y^*, \mathbf{X}, \mathbf{B}) \propto trN(m_0, M_0) \prod_{i=1}^n N(\mathbf{x}'_i \gamma_2 + T_i \beta + \theta_2 (T^*_i - \mathbf{x}'_i \gamma_1 - z_i \pi), \theta_1)
\]

where the proposal density is a non-central t-distribution centered at the mode of \( \prod_{i=1}^n N(\mathbf{x}'_i \gamma_2 + T_i \beta + \theta_2 (T^*_i - \mathbf{x}'_i \gamma_1 - z_i \pi), \theta_1) \) and the degrees of freedom set at some arbitrary value like 5.

**Calculating treatment effects**

In Equation (14), we show the formula for calculating the ATE in the Bayesian IV model with continuous outcomes. Under assumed constant effects, this is also equivalent to the TOT and LATE. However, it is also possible to generate posteriors for treatment effects with non-linear joint distributions for the treatment and the outcome.

In the case of the bivariate probit framework, the treatment effects are assumed to be heterogeneous due to the assumed non-linearity. We first estimate the interim quantity \( \Delta_{ATE,i} \) as

\[
\Delta_{ATE,i} = \phi(\mathbf{x}'_i \gamma_2 + \beta) - \phi(\mathbf{x}'_i \gamma_2)
\]
We evaluate this quantity at the observed value of the covariates $x_i$ to estimate the ATE, so

$$ATE = \frac{\sum_{i=1}^{n} \Delta_{ATE,i}}{n}$$

Calculating this value for each draw from the posterior provides a posterior distribution for ATE.

Another quantity of some interest is the effect of the treatment on the treated, defined as $TOT = E[y_1|T = 1] - E[y_0|T = 1]$. We can make use of the bivariate normal cumulative distribution function, which we denote $BN(.,.)$ to calculate a posterior for this value. We first calculate the interim value $\Delta_{TOT,i}$

$$\Delta_{TOT,i} = \frac{BN \left( \begin{pmatrix} x_i \gamma_1 + z_i \pi \\ x_i \gamma_2 + \beta \end{pmatrix}, \Omega \right) - BN \left( \begin{pmatrix} x_i \gamma_1 + z_i \pi \\ x_i \gamma_2 \end{pmatrix}, \Omega \right)}{\Phi(x'_i \gamma_1 + z_i \pi)}$$

again evaluated for the observed values of $x_i$ and $z_i$. The value of $TOT$ can be calculated as

$$TOT = \frac{\sum_{i=1}^{n} \Delta_{TOT,i}}{n}.$$

We can again calculate this value for each iteration to obtain a full posterior estimate.

A third quantity of interest is the local average treatment effect (LATE), which is interpreted as the effect of the treatment on the subpopulation that complies with the instrument (W. and Angrist 1994). Formally this is defined as

$$LATE = \frac{E[y|z = 1] - E[y|z = 0]}{E[T|z = 1] - E[T|z = 0]}$$

25 Here, we assume that the treatment effects are being calculated using the observed distribution of covariates. It would also be possible to calculate “predictive” treatment effects for some unknown distribution with specified values for covariates, instruments, and treatments.
For each individual, this is equivalent to

\[
\frac{\Pr(y_i = 1, T_i = 1 | z_i = 1) + \Pr(y_i = 1, T_i = 0 | z_i = 1) - \left[ \Pr(y_i = 1, T_i = 1 | z_i = 0) + \Pr(y_i = 1, T_i = 0 | z_i = 0) \right]}{\Pr(T_i = 1 | z = 1) - \Pr(T_i = 1 | z = 0)}.
\]

Let \(-\Omega\) be equivalent to \(\Omega\) except that the off-diagonal elements are multiplied by \(-1\). We can define the interim quantity

\[
\Delta_{LATE,i} = \frac{(M + N) - (O + P)}{\Phi(x_i' \gamma_1 + \pi) - \Phi(x_i' \gamma_1)}
\]

where

\[
M = BN \left( \begin{pmatrix} x_i \gamma_1 + \pi \\ x_i \gamma_2 + \beta \end{pmatrix}, \Omega \right)
\]

\[
N = BN \left( \begin{pmatrix} -(x_i \gamma_1 + \pi) \\ x_i \gamma_2 \end{pmatrix}, -\Omega \right)
\]

\[
O = BN \left( \begin{pmatrix} x_i \gamma_1 \\ x_i \gamma_2 + \beta \end{pmatrix}, \Omega \right)
\]

\[
P = BN \left( \begin{pmatrix} x_i \gamma_1 \\ -x_i \gamma_2 \end{pmatrix}, -\Omega \right)
\]

We evaluate this formula at the observed value of the covariates \(x\) for each observation \(i\). The value of LATE for each iteration of the MCMC sampler can be calculated as

\[
LATE = \frac{\sum_{i=1}^{n} \Delta_{LATE,i}}{n}
\]
APPENDIX C: ROBUSTNESS CHECKS FOR TABLE 5

To assess the robustness of our results, we re-estimated the results in Table 5 changing the prior distributions of the regression coefficients. Specifically, we choose prior distributions centered around zero, but with larger variances than those reported in the table. In Table 7, we show that the estimated effects and instrument strength remain unchanged despite the vague prior distributions.
Table 7: Bayesian IV robustness checks for the downstream effects of educational attainment on turnout

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>IHAD Bayesian, $\sum_{B_0} = I_k + 10$</th>
<th>ML</th>
<th>IHAD Bayesian, $\sum_{B_0} = I_k + 25$</th>
<th>ML</th>
<th>IHAD Bayesian, $\sum_{B_0} = I_k + 100$</th>
<th>ML</th>
<th>STAR Bayesian, $\sum_{B_0} = I_k + 10$</th>
<th>ML</th>
<th>STAR Bayesian, $\sum_{B_0} = I_k + 25$</th>
<th>ML</th>
<th>STAR Bayesian, $\sum_{B_0} = I_k + 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HS Graduation Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.140</td>
<td>-0.130</td>
<td>-0.127</td>
<td>-0.126</td>
<td>(0.295)</td>
<td>(0.303)</td>
<td>0.308</td>
<td>0.310</td>
<td>1.037</td>
<td>1.033</td>
<td>1.037</td>
<td>1.037</td>
</tr>
<tr>
<td></td>
<td>(0.158)</td>
<td>(0.158)</td>
<td>(0.162)</td>
<td>(0.158)</td>
<td>(0.204)</td>
<td>(0.204)</td>
<td>(0.202)</td>
<td>(0.202)</td>
<td>(0.063)</td>
<td>(0.652)</td>
<td>(0.065)</td>
<td>(0.065)</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.529</td>
<td>0.505</td>
<td>0.501</td>
<td>0.503</td>
<td>0.513</td>
<td>0.497</td>
<td>0.500</td>
<td>0.512</td>
<td>0.246</td>
<td>0.258</td>
<td>0.254</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>(0.228)</td>
<td>(0.229)</td>
<td>(0.230)</td>
<td>(0.229)</td>
<td>(0.383)</td>
<td>(0.376)</td>
<td>(0.378)</td>
<td>(0.376)</td>
<td>(0.118)</td>
<td>(0.124)</td>
<td>(0.129)</td>
<td>(0.124)</td>
</tr>
<tr>
<td><strong>Voter Turnout Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.602</td>
<td>-1.368</td>
<td>-1.364</td>
<td>-1.381</td>
<td>-1.013</td>
<td>-0.425</td>
<td>-0.432</td>
<td>-0.434</td>
<td>-1.850</td>
<td>-1.006</td>
<td>-0.874</td>
<td>-1.001</td>
</tr>
<tr>
<td></td>
<td>(0.553)</td>
<td>(0.397)</td>
<td>(0.418)</td>
<td>(0.410)</td>
<td>(0.752)</td>
<td>(0.529)</td>
<td>(0.549)</td>
<td>(0.545)</td>
<td>(0.778)</td>
<td>(0.565)</td>
<td>(0.658)</td>
<td>(0.591)</td>
</tr>
<tr>
<td>HS Graduation</td>
<td>1.084</td>
<td>0.660</td>
<td>0.647</td>
<td>0.683</td>
<td>1.051</td>
<td>0.132</td>
<td>0.143</td>
<td>0.153</td>
<td>1.948</td>
<td>0.970</td>
<td>0.826</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td>(0.888)</td>
<td>(0.660)</td>
<td>(0.688)</td>
<td>(0.686)</td>
<td>(1.070)</td>
<td>(0.747)</td>
<td>(0.774)</td>
<td>(0.766)</td>
<td>(0.888)</td>
<td>(0.652)</td>
<td>(0.751)</td>
<td>(0.680)</td>
</tr>
<tr>
<td>N</td>
<td>123</td>
<td>123</td>
<td>123</td>
<td>123</td>
<td>58</td>
<td>58</td>
<td>58</td>
<td>58</td>
<td>811</td>
<td>811</td>
<td>811</td>
<td>811</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.095</td>
<td>-0.084</td>
<td>-0.103</td>
<td>-0.103</td>
<td>-0.104</td>
<td>-0.110</td>
<td>-0.120</td>
<td>-0.218</td>
<td>-0.142</td>
<td>-0.219</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.387)</td>
<td>(0.396)</td>
<td>(0.400)</td>
<td>(0.431)</td>
<td>(0.438)</td>
<td>(0.444)</td>
<td>(0.370)</td>
<td>(0.422)</td>
<td>(0.391)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATE</td>
<td>0.148</td>
<td>0.144</td>
<td>0.153</td>
<td>0.153</td>
<td>0.137</td>
<td>0.040</td>
<td>0.043</td>
<td>0.296</td>
<td>0.249</td>
<td>0.293</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.160)</td>
<td>(0.168)</td>
<td>(0.167)</td>
<td>(0.167)</td>
<td>(0.262)</td>
<td>(0.271)</td>
<td>(0.268)</td>
<td>(0.187)</td>
<td>(0.227)</td>
<td>(0.195)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOT</td>
<td>0.110</td>
<td>0.104</td>
<td>0.112</td>
<td>0.112</td>
<td>0.025</td>
<td>0.028</td>
<td>0.031</td>
<td>0.288</td>
<td>0.241</td>
<td>0.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.143)</td>
<td>(0.156)</td>
<td>(0.145)</td>
<td>(0.145)</td>
<td>(0.265)</td>
<td>(0.273)</td>
<td>(0.270)</td>
<td>(0.181)</td>
<td>(0.224)</td>
<td>(0.190)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>