What Can Be Learned from a Simple Table? Bayesian Inference and Sensitivity Analysis for Causal Effects from 2×2 and $2 \times 2 \times K$ Tables in the Presence of Unmeasured Confounding*

Luke Keele† Kevin M. Quinn‡

March 28, 2014

Abstract

Presidents often campaign on behalf of Members of Congress during elections. Campaigning for co-partisan candidates may help that candidate win, which increases the probability that the president's party may have a majority in Congress. Due to powerful selection effects and unobserved confounding, it is difficult to estimate the effectiveness of these president campaign visits. The typical strategy to this empirical question would be to adjust for as many confounding variables as possible. Instead we ask what, if anything, should one infer about the causal effect of a presidential campaign visit using a simple cross-tabulation of of the data? Because the complete data likelihood under arbitrary patterns of confounding factorizes in a particularly convenient way, it is possible to parameterize this general situation with four easily interpretable parameters. Subjective beliefs regarding these parameters are easily elicited and subjective statements of uncertainty become possible. This paper also develops a novel graphical display called the *confounding plot* that quickly and efficiently communicates *all* patterns of confounding that would leave a particular causal inference relatively unchanged. Using data from the 2002 midterm elections, we find that under a reasonable set of assumptions presidential campaigning appears to have helped candidates win elections.

*An earlier version of this paper was awarded the 2008 Gosnell Prize for Excellence in Political Methodology. We thank Adam Glynn, Dan Ho, Gary King, and seminar participants at Penn State University and the Northeast Methods Meeting for helpful comments and conversations. Kevin Quinn gratefully acknowledges support from the U.S. National Science Foundation (grant BCS 05-27513) and the Institute for Quantitative Social Science at Harvard University. Software to perform the analyses described in this paper is available at http://cran.r-project.org/web/packages/SimpleTable/index.html.

†Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 Phone: 814-863-1592, Email: ljk20@psu.edu

‡Professor, UC Berkeley School of Law, 490 Simon #7200, University of California, Berkeley, Berkeley, CA 94720 Phone: 510-642-2485, Email: kquinn@law.berkeley.edu

1 Introduction: Example, Notation, and Elementary Analyses

1.1 Presidential Campaigning for Co-Partisans

The powers of U.S. presidents are largely informal. The veto and power of appointment are rare instances of formal powers given the president in the Constitution. Presidential power arises from the more informal power to persuade Congress (Neustadt 1960). One powerful means of persuasion is to make Members of Congress indebted to the president. One method that presidents have at their disposal to develop such indebtedness is the Congressional campaign visit. Here, the president makes a personal visit on behalf of someone running for the House or the Senate. Campaign visits by presidents are likely to be more effective for candidates to the House, which are often low information affairs. Voters may know little about a candidate for the U.S. House, so a presidential visit may serve as a powerful cue about the candidate given the amount of information available about the president.

Political scientists have attempted to estimate the causal effect of a presidential visit (Herrnson and Morris 2007; Sellers and Denton 2006; Cohen et al. 1991; Keele et al. 2004). Particular attention has been paid to the 2002 midterm election, when George W. Bush campaigned extensively for Republican candidates. A number of media reports inferred that a campaign visit that year was highly effective in helping candidates win (Keele et al. 2004). Of course, these analyses must rely on observational data, and the resulting estimates of causal effects are subject to confounding. To increase the credibility of the estimates as causal effects, analyst adjust for observed covariates. However, much of the data available are simply descriptive characteristics of the Congressional district such as the level of education or income in the district. Since the actual decision-making process of presidents is unobserved, there is little reason to believe that adjusted estimates are credible causal effects. Like many areas of study in the social sciences, experimentation is not possible, and estimated statistical

associations are in all likelihood highly confounded.

Many researchers might say little about causal effects can be learned from such data. While there are certainly aspects of truth to the position that unmeasured confounding simply cannot be overcome in many cases, it is possible to learn some things about the size of the causal effects of interest regardless of the nature and degree of the unmeasured confounding. For instance, Manski (1990) derived bounds for the average treatment effect under very general assumptions and showed that, in the case of binary treatment and outcome, the width of these bounds is 1.[1] Since the average treatment effect can take values between -1 and 1, Manski's bounding interval always includes the null value of no effect. Manski (2003) has also shown how auxiliary assumptions can narrow this bounding interval.

This paper further explores what can be learned about treatment effects under arbitrary forms of unmeasured confounding. We reduce the case of arbitrary unmeasured confounding with binary treatment and binary outcome to its most basic, yet still general, form that retains a readily interpretable parameterization of the key quantities. We then show how a Bayesian prior distribution can be placed over the four free parameters that govern the type and extent of the unmeasured confounding. If one is willing and able to use background knowledge to make some (possibly weak) assumptions about the nature of the unmeasured confounding, sharp posterior estimates of causal effects are easy to calculate. Since these assumptions are formalized within the Bayesian framework, subjective uncertainty about causal effects is calculated in a logically coherent manner. The end result is a procedure that allows researchers to make probability statements about the likely size of causal effects based on the evidence in a 2×2 (or $2 \times 2 \times K$) table regardless of the sample size and the amount of unmeasured confounding.

[1] For related work see Robins (1989); Manski (1993, 2003); Imai and Yamamoto (2008) and Balke and Pearl (1997).

1.2 Notation and Causal Model

Next, we present our notation and place our analysis in the formal statistical framework of causal inference based on potential outcomes (Holland 1986a). For each Congressional district $i = 1, 2, \ldots, 435$, we define two potential outcomes $Y_i(1)$ and $Y_i(0) \in [0, 1]$. $Y_i(1)$ denotes a potential electoral victory by the candidate in district i wins the election when the president campaigns for that candidate through a personal visit to that district. In contrast, $Y_i(0)$ represents a potential victory by the candidate in district i when the president does not campaign on the candidates behalf. We use the indicator variable $D_i \in [0, 1]$ to denote the treatment status in district i. In our application, D_i is equal to 1 if the president campaigned for candidate in district i through a personal visit and 0 otherwise. Under this causal model, the observed outcome is a function of the treatment variable and the potential outcomes: $Y_i = Y_i(1)D_i + (1 - D_i)Y_i(0)$. In our notation, we use upper-case letters to distinguish a random variable from its realization. The fundamental problem of causal inference is that for a single unit at most only one of the potential outcomes can be revealed (Holland 1986a).

This framework implicitly assumes that there is no interference among units: the potential units of one unit do not depend on the treatment of other units (Cox 1958; Rubin 1990). In our application, this assumption implies that the potential electoral victory status of a candidate in one district do not depend on whether the president campaigned for a candidate in another district. This assumption is reasonable given that presidents campaign for specific candidates and a visit is unlikely to help other candidates that did not specifically receive a presidential campaign visit.

Under this framework, the individual level causal effect is defined as a contrast in potential outcomes: $Y_i(1) - Y_i(0)$. Rather than focusing on unit-level causal effects, we will concern ourselves with aggregate effects within some collection of units. The causal quantity that we focus on is the average treatment effect (ATE)

$$ATE = E[Y_i(1) - Y_i(0)] \tag{1}$$

though many other causal estimands are possible such as the average treatment effect on the treated. This causal estimand, however, depends on counterfactual quantities. While we an easily calculate $E[Y_i = y|D_i = d]$ from the observed joint distribution of Y_i and D_i and use those in place of counterfactual quantities in Equation 1, these counterfactual quantities can only be estimated consistently from the joint distribution of the data if untestable causal assumptions are maintained (Rubin 1978; Holland 1986b; Robins 1986; Pearl 1995, 2000). Under SUTVA, if one is willing to assume that treatment assignment is strongly ignorable $Y_i(D) \perp\!\!\!\perp D_i$ then the estimated quantities will be equivalent to the counterfactual quantities in Equation 1. In a well-run randomized controlled experiment, strong ignorability of treatment assignment and SUTVA are likely to hold because of the design of the experiment.

However, with observational data, analyst often invoke the assumption of conditionally ignorability. Under conditional ignobility the claim is that there exists a collection of pre-treatment variables $\mathbf{U}$ such that treatment assignment is conditionally ignorable given $\mathbf{U}$.[2] Conditional ignorability is generally considered to be a strong assumption, since the analysts must assume that $\mathbf{U}$ contains all common causes of D_i and Y_i. Moreover, this assumption is not testable with data. We focus on what can be learned about Equation 1 when $\mathbf{U}$ is completely unobserved.

1.3 Data and Elementary Analyses

Our data is from the 2002 midterm election in the United States. These data were first reported in Keele et al. (2004). The outcome is measured as a binary indicator for whether the Republican candidate won the election or not and was constructed from *Congressional Quarterly's Politics in America* (2001). The treatment indicator is whether George W. Bush campaigned on behalf of U.S. House candidate through a personal appearance in that members district. Keele et al. (2004) used Lexis-Nexus state level AP reports from the 2002 election cycle to determine whether President George W. Bush personally campaigned for the Republican candidate between Labor Day and the election in November. The observed

[2]Here strong ignorability of treatment assignment holds within each level of $\mathbf{U}$.

data can be easily summarized in a 2×2 cross-tabulation as in Table 1. The cross-tabulation excludes races where one candidate ran unopposed.

Table 1: Observed Data For Presidential Visits in 2002

	$Y_i = 0$ Republican Loses	$Y_i = 1$ Republican Wins
No Visit $D_i = 0$	164	163
Visit $D_i = 1$	3	18

We see that the president visited 21 different Republican candidates. Of the 348 races with two candidates, the president only selected approximately six percent for a campaign visit. A naive analysis that assumes there is no confounding would clearly conclude that presidential visits are effective in helping Republican candidates win elections. The proportion of candidates that won when the president visited was .86 while the proportion of candidates that won without a presidential visit was .50, from these observed data quantities we can calculate the average treatment effect at .36 with a large sample 95% confidence interval (0.20, 1.0). Of course, confounding is likely as president's may strategically select candidates for visits when they perceive those candidates as being particularly able to win. We ask what can be learned about the ATE from Table 1 by placing reasonable assumptions on $\mathbf{U}$?

The paper proceeds as follows. In Section 2 we introduce the necessary terminology and notation and demonstrate how inferences can be constructed from a 2×2 table with general unmeasured confounding. Section 3 shows how causal quantities of interest such as the average treatment effect can be written in terms of the model parameters from Section 2. Large sample nonparametric bounds on these causal quantities are also derived in this section. These bounds coincide with those of Manski (1990) although the derivation is slightly different. Section 4 discusses the choice of prior distribution for the model parameters and

$Y_i(D_i = 0)$	$Y_i(D_i = 1)$	Z_i	
0	0	0	Never Succeed
0	1	1	Helped
1	0	2	Hurt
1	1	3	Always Succeed

Table 2: *Possible Patterns of Potential Outcomes and Coarsest General Confounding Variable.* A unit i for which $Z_i = 0$ has a value of U_i that causes it to always have $Y_i = 0$ regardless of the (counterfactual) value of D_i. We say these units are "never succeeders". If $Z_i = 1$ we say that unit i is "helped" by treatment because its potential outcome under $D_i = 1$ is equal to 1 (success) while its potential outcome under $D_i = 0$ is 0 (failure). If unit i has $Z_i = 2$ we say that i is "hurt" by treatment because its potential outcome under $D - i = 1$ is equal to 0 (failure) while its potential outcome under $D_i = 0$ is 1 (success). Finally, if $Z_i = 3$ we say that i is an "always succeeder" because its value of U_i is such that Y_i will always equal 1 regardless of the (counterfactual) value of D_i.

then describes a simple posterior sampling algorithm that does not require Markov chain Monte Carlo. Section 4 also describes the construction and interpretation of the novel confounding plot discussed above. In Section 5 we revisit the example data in Table 1. Here we see how defensible prior beliefs can be operationalized in a prior distribution over the model parameters and what this implies for inferences about a possible presidential visit treatment effect. The final section concludes.

2 Probability Model under Unobserved Confounding

While $\mathbf{U}$ may be extremely complicated, the binary nature of both treatment and outcome implies that the domain of $\mathbf{U}$ can be partitioned into four equivalence classes depending on the pattern of potential outcomes associated with each point in the domain of $\mathbf{U}$ (Angrist et al. 1996; Balke and Pearl 1997; Chickering and Pearl 1997). We introduce a new categorical variable Z_i that labels these equivalence classes. The values of Z_i along with the associated patterns of potential outcomes are presented in Table 2.

If Z_i were observed, one could write the post-intervention distribution, P_{DYZ}, as

$$\Pr(Y(D = d) = y) = \sum_{z=0}^{3} \Pr(Y = y | D = d, Z = z) \Pr(Z = z).$$

The probabilities on the right-hand-side of the equation above can be calculated directly from P_{DYZ}. In our application, as is common in the social sciences, Z_i is unobservable for any i. Without data on Z_i, it is impossible to consistently estimate P_{DYZ}. Nevertheless, there is some information about Z_i in observed (D, Y) data sampled from P_{DYZ}. The goal of this paper is to show how this information can be combined with subjective background knowledge to yield causal inferences from 2×2 and $2\times2\times K$ tables even when the confounding variables in $\mathbf{U}$ are not measured.

2.1 Likelihood

We adopt a Bayesian approach to make inferences about the form of P_{DYZ} without information on Z_i. The main reason for taking a Bayesian approach in this paper is that it allows us to incorporate background knowledge about the (potentially unobserved) confounder Z_i in a principled fashion (Kadane and Wolfson 1998; Western and Jackman 1994; Gill and Walker 2005). We begin by discussing the likelihood function and then discuss our choice of prior distribution along with the resulting posterior distribution.

Let $\mathcal{Z}_i$ denote the set of possible values Z_i could take given the observed data on unit i. More formally,

$$\mathcal{Z}_i = \begin{cases} \{0,1\} & \text{if } d_i = 0,\ y_i = 0 \\ \{2,3\} & \text{if } d_i = 0,\ y_i = 1 \\ \{0,2\} & \text{if } d_i = 1,\ y_i = 0 \\ \{1,3\} & \text{if } d_i = 1,\ y_i = 1 \end{cases}$$

We can then write the likelihood as:

$$
\begin{aligned}
p(\mathbf{d}, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) &= \prod_{i=1}^{n} \sum_{z_i \in \mathcal{Z}_i} p(d_i, y_i, z_i|\boldsymbol{\theta}, \boldsymbol{\psi}) \\
&= \prod_{i=1}^{n} p(d_i, y_i|\boldsymbol{\theta}) \left\{ \sum_{z_i \in \mathcal{Z}_i} p(z_i|d_i, y_i, \boldsymbol{\psi}) \right\} \\
&= \prod_{i=1}^{n} \theta_{00}^{\mathbb{I}(d_i=0,y_i=0)} \theta_{01}^{\mathbb{I}(d_i=0,y_i=1)} \theta_{10}^{\mathbb{I}(d_i=1,y_i=0)} \theta_{11}^{\mathbb{I}(d_i=1,y_i=1)} \times \\
&\qquad \left\{ \sum_{z_i \in \mathcal{Z}_i} \psi_{00}^{\mathbb{I}(d_i=0,y_i=0,z_i=1)} (1-\psi_{00})^{\mathbb{I}(d_i=0,y_i=0,z_i=0)} \times \right. \\
&\qquad\quad \psi_{01}^{\mathbb{I}(d_i=0,y_i=1,z_i=3)} (1-\psi_{01})^{\mathbb{I}(d_i=0,y_i=1,z_i=2)} \times \\
&\qquad\quad \psi_{10}^{\mathbb{I}(d_i=1,y_i=0,z_i=2)} (1-\psi_{10})^{\mathbb{I}(d_i=1,y_i=0,z_i=0)} \times \\
&\qquad\quad \left. \psi_{11}^{\mathbb{I}(d_i=1,y_i=1,z_i=3)} (1-\psi_{11})^{\mathbb{I}(d_i=1,y_i=1,z_i=1)} \right\} \\
&= \theta_{00}^{C_{00+}} \theta_{01}^{C_{00+}} \theta_{10}^{C_{10+}} \theta_{11}^{C_{11+}} \qquad (2)
\end{aligned}
$$

where $\mathbb{I}(\cdot)$ is the indicator function, $C_{dy+} = \sum_{i=1}^{n} \mathbb{I}(d_i = d, y_i = y)$, $\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11} \geq 0$, $\theta_{00} + \theta_{01} + \theta_{10} + \theta_{11} = 1$, and $\psi_{00}, \psi_{01}, \psi_{10}, \psi_{11} \in [0, 1]$.

While this model for (D, Y, Z) might seem to contain a large number of parameters, there are two key sets of parameters θ_{dy} and ψ_{dy} for $d = 0, 1$ and $y = 0, 1$. The θ parameters govern a multinomial distribution for the distribution of (D_i, Y_i) after Z_i has been marginalized out of P_{DYZ}. The ψ parameters govern the conditional distribution of Z_i given D_i and Y_i. Note that because of the definition of Z_i (see Table 2) only 2 values of Z are logically possible given any admissible (D_i, Y_i) pair. The distribution of Z_i given $D_i = d$ and $Y_i = y$ is thus Bernoulli with parameter ψ_{dy}. Here, ψ_{01} gives the probability that $Z_i = 3$ given $D_i = 0$ and $Y_i = 1$ while $(1 - \psi_{01})$ gives the probability that $Z_i = 2$ given $D_i = 0$ and $Y_i = 1$. The other conditional distributions for Z_i given $D_i = d$ and $Y_i = y$ are similarly parameterized. Table 6 in the appendix provides a complete summary of the parameters and their intuitive meanings.

2.2 Prior and Posterior

Bayesian inference centers on the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ given the observed data. The posterior distribution is given (up to proportionality) by:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, \mathbf{d}, \mathbf{z}) \propto p(\mathbf{y}, \mathbf{d}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi})$$

We defined the likelihood, $p(\mathbf{y}, \mathbf{d}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\psi})$, in the previous section, and we now discuss specification of the prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\psi})$. A natural choice for the joint prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is to assume that $\boldsymbol{\theta}$, $\psi_{00}, \psi_{01}, \psi_{10}$, and ψ_{11} are mutually independent a priori and that $\boldsymbol{\theta} \sim \mathcal{D}irichlet(a_{00}, a_{01}, a_{10}, a_{11})$, $\psi_{dy} \sim \mathcal{B}eta(b_{dy}, c_{dy})$, for $d = 0, 1$ and $y = 0, 1$. This is the conjugate prior distribution for this model. This prior specification will allow us to think of the hyper-parameters a_{dy}, b_{dy}, and c_{dy} for $d = 0, 1$ and $y = 0, 1$ as additional "pseudo-observations." This makes the prior distributions more easily interpretable, which is important for the current application where inferences are dependent on the prior.

Combining this prior with the likelihood in Equation 2 gives us the following posterior density (up to proportionality):

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{d}, \mathbf{y}) \propto \theta_{00}^{C_{00+}+a_{00}-1} \theta_{01}^{C_{01+}+a_{01}-1} \theta_{10}^{C_{10+}+a_{10}-1} \theta_{11}^{C_{11+}+a_{11}-1} \times \\ \psi_{00}^{b_{00}-1}(1-\psi_{00})^{c_{00}-1} \psi_{01}^{b_{01}-1}(1-\psi_{01})^{c_{01}-1} \times \\ \psi_{10}^{b_{10}-1}(1-\psi_{10})^{c_{10}-1} \psi_{11}^{b_{11}-1}(1-\psi_{11})^{c_{11}-1} \end{aligned} \tag{3}$$

Note that the only information about $\boldsymbol{\psi}$ is coming from the prior distribution. This implies that inferences that depend on $\boldsymbol{\psi}$ will be dependent on one's choice of prior for $\boldsymbol{\psi}$.

3 Causal Quantities of Interest

Next we discuss how causal quantities such as the ATE can be calculated based on beliefs about $\boldsymbol{\psi}$. Typically causal quantities are calculated directly from the data based on an as-

sumption of conditional unconfoundedness. In our analysis of whether presidential campaign visits help candidates, unconfoundness, conditional or otherwise, is unrealistic. In our approach, based on the probability model specified in Section 2, we calculate causal quantities that depend on the parameter $\boldsymbol{\psi}$. We denote these as sensitivity analysis quantities since the inference is based on prior beliefs about the unobserved distribution of Z_i. Sensitivity analysis quantities depend on the distribution of Z_i through the following set of equations:

$$\begin{aligned}
\Pr_s(Y(D=0)=0) &= \sum_{z=0}^{3} \Pr(Y=0|D=0,Z=z)\Pr(Z=z) \\
&= \Pr(Z=0)+\Pr(Z=1) \\
&= \theta_{10}(1-\psi_{10})+\theta_{11}(1-\psi_{11})+\theta_{00} \\
\Pr_s(Y(D=0)=1) &= \sum_{z=0}^{3} \Pr(Y=1|D=0,Z=z)\Pr(Z=z) \\
&= \Pr(Z=2)+\Pr(Z=3) \\
&= \theta_{10}\psi_{10}+\theta_{11}\psi_{11}+\theta_{01} \\
\Pr_s(Y(D=1)=0) &= \sum_{z=0}^{3} \Pr(Y=0|D=1,Z=z)\Pr(Z=z) \\
&= \Pr(Z=0)+\Pr(Z=2) \\
&= \theta_{00}(1-\psi_{00})+\theta_{01}(1-\psi_{01})+\theta_{10} \\
\Pr_s(Y(D=1)=1) &= \sum_{z=0}^{3} \Pr(Y=1|D=1,Z=z)\Pr(Z=z) \\
&= \Pr(Z=1)+\Pr(Z=3) \\
&= \theta_{00}\psi_{00}+\theta_{01}\psi_{01}+\theta_{11}.
\end{aligned}$$

If $\boldsymbol{\psi}$ were known the sensitivity analysis post-intervention distribution would yield the true post intervention distribution. Of course, $\boldsymbol{\psi}$ is never known (and typically not identified) so the sensitivity analysis post-intervention distribution will depend on one's prior beliefs

about $\boldsymbol{\psi}$.

3.1 Average Treatment Effects

Next, we describe how causal quantities can be calculated based on values of $\boldsymbol{\psi}$. Here, we focus on the ATE. While we focus on the ATE, it is also possible to define average treatment effects on the treated (ATT) or just the control group (ATC) and calculate bounds and sensitivity analysis distributions for these estimands as well. Moreover, we can also calculate sensitivity analysis quantities based on the relative risk as well. The sensitivity analysis ATE is defined as:

$$\begin{aligned} ATE_s &= \Pr_s(Y(X=1)=1) - \Pr_s(Y(X=0)=1) \\ &= (\theta_{00}\psi_{00} + \theta_{01}\psi_{01} + \theta_{11}) - (\theta_{10}\psi_{10} + \theta_{11}\psi_{11} + \theta_{01}) \end{aligned} \tag{4}$$

Manski (1990) derived nonparametric bounds for the average treatment effect that will contain the true average treatment effect with probability 1 as sample size goes to infinity. Here, we show how these bounds can be calculated as a function $\boldsymbol{\psi}$. Inspection of Equation 4 reveals that the minimum value of ATE_s will occur when $\psi_{00} = 0, \psi_{01} = 0, \psi_{10} = 1$, and $\psi_{11} = 1$. Similarly, the maximum value of ATE_s will occur when $\psi_{00} = 1, \psi_{01} = 1, \psi_{10} = 0$, and $\psi_{11} = 0$. Substituting these values into the expression for ATE_s and recognizing that $ATE_s = ATE$ we see that:

$$ATE \in [-(\theta_{10} + \theta_{01}), \quad (\theta_{00} + \theta_{11})]$$

Substituting the MLEs for $\theta_{00}, \theta_{01}, \theta_{10}$, and θ_{11} we see that (in a slight abuse of notation)

$$\lim_{n\to\infty} \Pr\left(\frac{-C_{01+} - C_{10+}}{n} \leq ATE \leq \frac{C_{00+} + C_{11+}}{n}\right) = 1$$

where it is understood that the C_{xy+} counts also depend on n. Note that this interval will always include 0. Further, as Manski (1990) has shown, and as is easy to see here since $\sum_x \sum_y \theta_{xy} = 1$, the width of this interval will always be 1.

4 Bayesian Inference For Causal Effects

We next describe a Bayesian method for obtaining sensitivity analysis quantities. Here, we focus on obtaining sensitivity analysis quantities for the ATE, but sensitivity analysis quantities are easily obtained for the ATT or ATC. A Bayesian approach requires one to specify a prior distribution for $(\boldsymbol{\theta}, \boldsymbol{\psi})$. In Section 2 we argued that independent Dirichlet and Beta distributions made sense in terms of interpretability. In the remainder of this section we discuss how the parameters governing these prior distributions can be chosen and how one can summarize the resulting posterior distribution to make inferences about causal quantities of interest such as the ATE under general but unobservable patterns of confounding.

4.1 Choosing a Prior Distribution

It is worth emphasizing that, unlike Bayesian inference for models which are point identified, the impact of the choice of the prior for $\boldsymbol{\psi}$ on the posterior distribution for $\boldsymbol{\psi}$ and functionals of that posterior distribution will not diminish as n gets large if Z is completely unobserved. In fact, since no new information about $\boldsymbol{\psi}$ is arriving as n gets large, the marginal posterior for $\boldsymbol{\psi}$ will always be equal to the prior for $\boldsymbol{\psi}$. Specifically, one should be able to justify a particular choice of prior by an appeal to substantive background knowledge. Moreover, the analyst should report numerous sensitivity analysis quantities in which multiple reasonable priors are used.

Each ψ_{xy} represents the conditional probability of one of the two possible configurations of potential outcomes among units in which we observe $X = x$ and $Y = y$.[3] Thus the

[3]See the appendix for a full elaboration of how each ψ_{xy} parameter relates to a particular set of potential outcomes.

$\mathcal{B}eta(b_{xy}, c_{xy})$ prior for ψ_{xy} can be thought of as a statement of belief that $b_{xy} - 1$ of the C_{xy+} units have one potential outcome profile while $c_{xy} - 1$ of the C_{xy+} units have the other possible potential outcome profile. If $b_{xy} + c_{xy} = C_{xy+} + 2$ then the information in the prior is equivalent to the information that would be in the sample data in the ideal case in which the potential outcome patterns are observed for units with $X = x$ and $Y = y$. If $b_{xy} + c_{xy} < C_{xy+} + 2$ then there is less information in the prior than this ideal situation and if $b_{xy} + c_{xy} > C_{xy+} + 2$ then the prior is adding more information than one could ever get directly from the sample data. In the appendix, we provide a full elaboration of the relationship between the b_{xy} and c_{xy} parameters and potential outcomes. Here, we provide two useful heuristics for prior selection based on two models of treatment response and selection. We use both of these heuristics in the analysis of the presidential visit data from 2002.

The first heuristic we consider is the possibility of a monotonic treatment effect (Manski 1997). Under monotonicity, we assume

$$Y_i(1) \geq Y_i(0) \text{ or } Y_i(1) \leq Y_i(0) \; \forall \; i = 1, \cdots, n.$$

In words, under monotonicity we assume that outcomes for the treated are greater than or no smaller (less than or no larger) than those in the control condition. In the context, here, the monotone treatment response assumption implies that a presidential campaign visit does not hurt the election chances of any candidates. If one believes that a generally positive monotonic treatment effect is reasonable then one could set $b_{01} \gg c_{01}$ and $b_{10} \ll c_{10}$. Conversely, one could set $b_{00} \ll c_{00}$ and $b_{11} \gg c_{11}$ to operationalize a generally negative monotonic treatment effect. Setting the parameters on the prior distributions as follows $b_{01} = \infty$, $c_{01} = 0$ and $b_{10} = 0$, $c_{10} = \infty$ operationalizes Manski's version of the monotone treatment response assumption. Manski's version of monotone treatment response is, in fact, a limiting case of monotonic treatment effects. We need not assume that treatment is

uniformly positive for all units. We can vary the fraction of units helped or hurt to assess sensitivity of the resulting estimates to the monotonicity assumption. Monotone treatment response is a fairly weak assumption in this context. In general, it seems unlikely that a campaign visit would hurt a candidates vote share even if the visit did not help the candidate's election chances.

The next heuristic we consider is that of treatment selection (Manski 1995). Under selection, we assume that treated units are selected to maximize the outcome. This implies that the president selects candidates for a visit based on whether the visit is likely to increase the chances of winning that election. Formally, we write a selection assumption in the following way:

$$Pr(\text{Helped}|D = 1) \geq Pr(\text{Hurt}|D = 0)$$

$$Pr(\text{Hurt}|D = 1) \geq Pr(\text{Helped}|D = 0)$$

To operationalize selection effects through the prior, one could set $c_{11}/(b_{11}+c_{11}) > b_{00}/(b_{00}+c_{00})$. If one believes that units avoid selecting a treatment that harms them one could set $c_{01}/(b_{01}+c_{01}) > b_{10}/(b_{10}+c_{10})$. Is selection a reasonable assumption in this context? The president is unlikely to waste time campaigning for candidates that have little chance of winning nor is the president likely to campaign much for candidates that will win easily. Thus the president is likely to avoid never succeed and always succeed types and instead attempt to identify candidates that will be helped. Finally, we can combine the monotonicity and selection assumptions and estimates effects under the assumption of monotone treatment selection (MTS), which should further sharpen the inference. While, we cannot verify that either assumption holds individually or in combination, there is no reason to think that the presence of one assumption decreases the likelihood of the other assumption. While researchers may choose different prior values, we think these two heuristics are widely applicable to many empirical settings.

4.2 Posterior Inference

The posterior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ discussed in Section 2 can all be sampled using simple independent Monte Carlo sampling. Markov chain Monte Carlo methods are unnecessary. To produce a Monte Carlo sample of size m from the distribution with density given (up to proportionality) by Equation 3 we can use Algorithm 4.1.

Algorithm 4.1: PosteriorSamplingUnobservedZ($\boldsymbol{C}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, m$)

for $j \leftarrow 1$ **to** m

do
$$\begin{cases} \boldsymbol{\theta}^{(j)} \leftarrow rdirichlet(C_{00+} + a_{00},\ C_{01+} + a_{01},\ C_{10+} + a_{10},\ C_{11+} + a_{11}) \\ \psi_{00}^{(j)} \leftarrow rbeta(b_{00},\ c_{00}) \\ \psi_{01}^{(j)} \leftarrow rbeta(b_{01},\ c_{01}) \\ \psi_{10}^{(j)} \leftarrow rbeta(b_{10},\ c_{10}) \\ \psi_{11}^{(j)} \leftarrow rbeta(b_{11},\ c_{11}) \end{cases}$$

return $(\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{m},\ \{\psi_{00}^{(j)}\}_{j=1}^{m},\ \{\psi_{01}^{(j)}\}_{j=1}^{m},\ \{\psi_{10}^{(j)}\}_{j=1}^{m},\ \{\psi_{11}^{(j)}\}_{j=1}^{m})$

Here $rdirichlet(d, e, f, g)$ is a function that returns a pseudo-random draw from a $\mathcal{D}irichlet(d, e, f, g)$ distribution and $rbeta(d, e)$ is a function that returns a pseudo-random draw from a $\mathcal{B}eta(d, e)$ distribution.

Once a sample $\{\boldsymbol{\theta}^{(j)}, \boldsymbol{\psi}^{(j)}\}_{j=1}^{m}$ from the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ has been drawn, these draws are plugged into the formulas for the causal quantity of interest. A sample from the posterior distribution of the sensitivity analysis average treatment effect ($\{ATE_s^{(j)}\}_{j=1}^{m}$) can be constructed by taking the jth sample to be

$$ATE_s^{(j)} = \left(\theta_{00}^{(j)}\psi_{00}^{(j)} + \theta_{01}^{(j)}\psi_{01}^{(j)} + \theta_{11}^{(j)}\right) - \left(\theta_{10}^{(j)}\psi_{10}^{(j)} + \theta_{11}^{(j)}\psi_{11}^{(j)} + \theta_{01}^{(j)}\right)$$

for $j = 1, \ldots, m$. Samples from the posterior distributions of other causal quantities of interest follow analogously. Once we obtain a sample from the posterior distribution of interest, we can summarize the distribution by calculating density estimates, highest posterior density

regions (the smallest region that contains a pre-specified amount of the posterior mass), the probability that a quantity of interest is greater than 0, etc. using the sampled parameter values. See Jackman (2000); King et al. (2000); Gelman et al. (2003) and Gill (2007) for discussions of how posterior samples can be summarized.

4.3 The Confounding Plot

While the subjective Bayesian approach outlined above takes beliefs about $\boldsymbol{\psi}$ as input and returns a subjective probability distribution over causal effects, we can also reverse this process and start with a particular prima facie post-intervention distribution and ask what values of $\boldsymbol{\psi}$ will result in a sensitivity analysis post-intervention distribution that is within some tolerance of the given prima facie post-intervention distribution.

Formally, for a given values of $\theta_{00}, \theta_{01}, \theta_{10}$ and θ_{11} we seek to find all values of $\psi_{00}, \psi_{01}, \psi_{10}$ and ψ_{11} for which

$$\begin{aligned} &|\Pr_p(Y(X=0)=0) - \Pr_s(Y(X=0)=0)| = \\ &\left| \frac{\theta_{00}}{\theta_{00}+\theta_{01}} - (\theta_{10}(1-\psi_{10}) + \theta_{11}(1-\psi_{11}) + \theta_{00}) \right| \leq \epsilon \end{aligned} \tag{5}$$

and

$$\begin{aligned} &|\Pr_p(Y(X=0)=1) - \Pr_s(Y(X=0)=1)| = \\ &\left| \frac{\theta_{01}}{\theta_{00}+\theta_{01}} - (\theta_{10}\psi_{10} + \theta_{11}\psi_{11} + \theta_{01}) \right| \leq \epsilon \end{aligned} \tag{6}$$

and

$$\begin{aligned} &|\Pr_p(Y(X=1)=0) - \Pr_s(Y(X=1)=0)| = \\ &\left| \frac{\theta_{10}}{\theta_{10}+\theta_{11}} - (\theta_{00}(1-\psi_{00}) + \theta_{01}(1-\psi_{01}) + \theta_{10}) \right| \leq \epsilon \end{aligned} \tag{7}$$

and

$$|\Pr_p(Y(X=1)=1) - \Pr_s(Y(X=1)=1)| = \left| \frac{\theta_{11}}{\theta_{10}+\theta_{11}} - (\theta_{00}\psi_{00} + \theta_{01}\psi_{01} + \theta_{11}) \right| \leq \epsilon \tag{8}$$

for some small positive ϵ. Note that Inequalities 5 and 6 only depend on ψ_{10} and ψ_{11} while Inequalities 7 and 8 only depend on ψ_{00} and ψ_{01}. It is thus possible to depict all values of $\psi_{00}, \psi_{01}, \psi_{10}$ and ψ_{11} that satisfy Inequalities 5 - 8 with a pair of 2-dimensional plots— one of ψ_{10} and ψ_{11} and another of ψ_{00} and ψ_{01}.

Because this method does not account for sampling variability it is most appropriate for situations in which all of the cells in the 2×2 table have a reasonably large number of observations. As such, for our application, the confounding plot is of limited usefulness since the samples sizes in the table are rather limited. We note in passing the similarity of the plots above to the tomography plots of King (1997) that are useful for ecological inference. Indeed, the situation under consideration in this paper in which C_{xy+} are all fully observed but the joint C_{xyz} counts are not observed can be thought of as a particular type of ecological inference problem (Richardson 2004).

5 Sensitivity analysis quantities for presidential campaigning

We begin the analysis with the estimation of naive treatment effect of a presidential campaign visit in 2002. For this estimate to be a valid causal effect, we must assume that presidential visits were as-if randomly assigned across Congressional Districts. An assumption that is clearly implausible. The naive treatment effect estimate is $18/(18+3) \approx .857) - (163/(164+163) \approx .498) \approx .36$. This estimate is clearly bound away from zero with a 95% confidence interval of 0.189 and 0.493. In contrast, the no-assumption bounds on the treatment effect, often called Manski bounds reported in Table 3 are -0.477 and 0.522. In a

sensitivity analysis, we hope to accomplish two goals. One goal is to produce an inference under assumptions that are more realistic than the naive estimate of the causal effect but also to sharpen the inference over the no assumption bounds, while using set of credible but weak set of assumptions.

Next, we use a series of prior distributions to in an attempt to achieve these two goals. We begin with an uninformative prior. Under the uninformative prior, we assume that the fraction of units helped by the treatment ($T = 1, Y = 0$) is equal to the fraction of the units that is helped by the treatment ($T = 0, Y = 0$). The estimate and 95% confidence intervals are in Table 3. Under this assumption, the bounds on the treatment effect are now -0.348 and 0.396, with a point estimate of 0.023. Unlike in many applications of Bayesian inference, uninformative priors make little sense, since the goal is to use substantive information to reason about the nature of confounding.

Table 3: Estimated Treatment Effects for Presidential Visits

Naive Estimate	No-Assumption Bounds	Uniform Prior
0.36		0.023
[0.189, 0.493]	[-0.477, 0.522]	[-0.348, 0.396]

Next, we use a set of priors that assume varying levels of a monotonic treatment effect. Here, we make assumptions about the about the fraction of units helped by the treatment. The monotonicity assumption seems reasonable; given that we expect that a visit from George W. Bush is unlikely to hurt the vote shares of few if any candidates. See the appendix for a full report of the prior values used in the analysis. The first row of Table 4 contains estimates for the effect of a presidential visit under three different levels of monotonicity. The strongest monotonicity assumption used here is nearly equivalent to Manski's(1997) definition of monotone treatment response. Under the weakest monotonicity assumption, the estimated treatment effect is 0.10, however the bounds on this estimate include zero and thus don't allow us to conclude that visits were effective. Strengthening the monotonicity

assumption increases the point estimate to nearly 0.22 much closer to the naive estimate, but again the bounds on this estimate include zero. Under the strongest monotonicity assumption, the inference is now bound away from zero with a point estimate of 0.261. This a sizable effect but still more than 25% smaller than the naive treatment effect estimate. It appears that so long as we are willing to assume that the effect of the treatment was strongly monotonic presidential visits were an effective campaign strategy. But for this to be a valid inference, we must be willing to believe that a visit by George W. Bush did not hurt the reelection chances of any candidates.

Table 4: Estimated Treatment Effects for Presidential Visits

Monotonicity		
Weak	Medium	Strong
0.102	0.219	0.261
[-0.161, 0.371]	[-0.025, 0.475]	[0.029, 0.491]
Selection		
Weak	Medium	Strong
0.052	0.081	0.119
[-0.213, 0.315]	[-0.189, 0.346]	[-0.141, 0.380]

Next, we use a set of priors that assume that candidates in the treatment group where chosen to maximize the outcome: wins by Republican candidates. Thus we assume that Bush chose to campaign for candidates that he thought would be most helped by a visit. Given that presidents have many constraints on their time, the selection assumption is quite plausible. We vary the strength of this assumption and estimate the treatment effect under weak, medium, and strong selection assumptions. The estimates under selection are in the second row of Table 4. The selection assumption improves the inference over the no-assumption bounds but not enough to make the inference informative. Under weak selection, the point estimate is consistent with only a small treatment effect at 0.05, and the bounds

for this estimate include zero. The two stronger selection assumptions increase the point estimates to 0.08 and 0.12 respectively. The bounds for both estimates, however, include zero. Therefore, the selection assumption at any level is not sufficient to allow us to conclude that presidential visits were an effective campaign tool in 2002.

Table 5: Estimated Treatment Effects for Presidential Visits

		Monotonicity		
		Weak	Medium	Strong
Selection	Weak	0.131 [-0.029, 0.288]	0.247 [0.111, 0.38]	0.289 [0.182, 0.398]
	Medium	0.161 [0.005, 0.314]	0.276 [0.139, 0.407]	0.310 [0.208, 0.428]
	Strong	0.199 [0.043, 0.348]	0.315 [0.181, 0.440]	0.361 [0.251, 0.459]

Finally, we use priors that assume both selection and monotonicity of varying levels. If we again vary each assumption at three levels, that creates nine different prior combinations. Table 5 contains the nine different estimates of the treatment effect along with the associated 95% credible intervals. The combination of the monotonicity and selection assumptions is sufficient to result in informative inferences in five out of the six combinations. Only under if treatment response is weakly monotonic and selection is weak does the credible interval contain zero. Under the other five combinations the treatment effect varies from 0.161 to 0.361, which matches the naive estimate of the treatment effect. Second, we see that magnitude of the treatment effect is mostly responsive to the monotonicity assumption. An increase in the level of selection increases the treatment effect point estimate by around 0.03 across all levels of monotonicity. We see the largest increase in the point estimate as we move from weak to a medium level of monotonicity.

What conclusion should we draw about the effectiveness of presidential campaign visits?

Our conclusions here must be evaluated against the plausibility of the assumptions. The selection assumption is the most likely of the two assumptions to hold. There is every reason to believe that presidents' select which candidates to campaign for based some hope of being effective. It is unlikely that Presidents would spend much time campaigning for candidates that have little chance of winning. As such, in this context the selection assumption is the most credible, but it also does not have much power. Based on the selection assumption alone we cannot rule out that the average treatment effect is zero.

If we believe that treatment response is monotone, however, there does appear to be a treatment effect. However, with only the monotonicity assumption, treatment is only effective if we believe that none of the candidates election chances were hurt by a presidential visit. This is probably implausible. However, if combine these two assumptions, however, the inference is informative. The most plausible combination of assumption is that of treatment responses that are moderately monotonic with medium to strong levels of selection. While selection probably holds, assuming that no candidates were hurt by visit is less plausible. Under such a set of assumptions, this would imply that the treatment effect is between 0.161 and 0.315. This implies that while presidential visits were effective, they were, however, less effective than one would find if we assume the visits were randomly assigned.

We can further explore the role of assumption using a confounding plot. In a confounding plot, we can infer what set of assumptions would produce an inference as found under the naive estimate of the treatment effect. In a confounding plot, we observe what fraction of units would be needed in each cell to produce the naive estimate of the treatment effect within some specified tolerance level for this data.

Figure 1 contains the confounding plot of the 2002 presidential visit data. The confounding plot contains the range of fractions necessary within each cell of Table 1 to produce the treatment effect estimate within 0.025 of the naive estimate in Table 3. I focus on the right hand plot in the Figure. In particular the right hand x-axis contains the range of what fraction of units in that cell had to be helped by a visit to produce an estimate near the naive

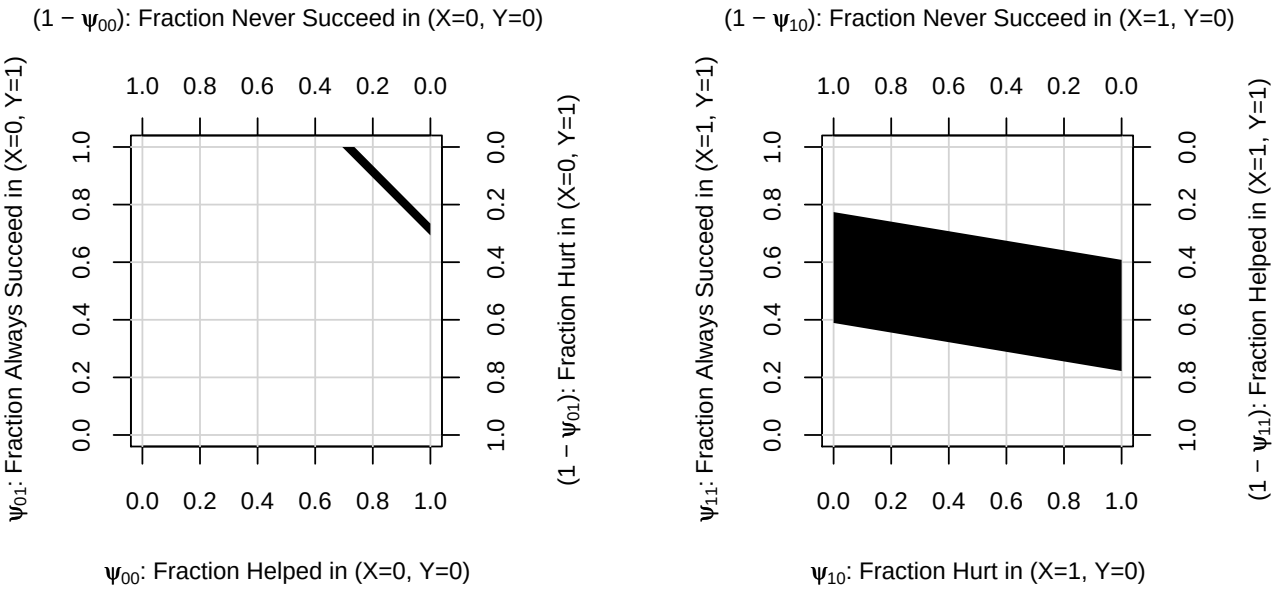

Figure 1: Confounding Plot for Treatment Effect of Presidential Campaign Visit, 2002

treatment estimate. Here we see that the fraction of helped individuals in the condition $T_i = 1, Y_i = 1$ must be quite high or above 0.80, while the fraction never succeed candidates must be quite high in condition $T_i = 0, Y_i = 0$. This highlights the importance of the monotonicity assumption in the earlier analyses. If the fraction of helped candidates is quite large then the monotonicity assumption holds, and it was only under strong monotonicity that we observed treatment effect estimates that were large and bounded away from zero. Thus this underscores the role of the monotonicity assumption in the analysis.

6 Conclusion

In this article we have illustrated how to conduct a form of sensitivity analysis under general patterns of unobserved confounding. In many social science applications interventions cannot be randomized, and the assumption of no confounding is is implausible. In our application, there is little reason to think we observe all the necessary covariates that would make congressional districts visited by the president comparable to the congressional districts that do

not receive a presidential campaign visit. The development of methods of sensitivity analyses for situations in which unmeasured confounding is present, as is done in this paper, serves to shift empirical social science research away from the all too typical enterprise of defending indefensible causal assumptions to the practice of honestly stating the range of assumptions that are consistent with a particular type of causal effect. Using Bayesian methods, analysts can present a range of estimates under different patterns of confounding. In the analysis presented, here, we find that under a monotonicity assumption it appears that presidential visit aided candidates, while assumptions about treatment selection did little to alter the inference about the presidential visit treatment effect.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Balke, A. and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Chickering, D. M. and Pearl, J. (1997), "A Clinician's Tool for Analyzing Non-Compliance," *Computing Science and Statistics*, 29, 424–431.

Cohen, J. E., Krassa, M. A., and Hamman, J. A. (1991), "The Impact of Presidential Campaigning on Midterm U.S. Senate Elections," *American Political Science Review*, 85, 165–178.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis*, London: Chapman & Hall, 2nd ed.

Gill, J. (2007), *Bayesian Methods: A Social and Behavioral Science Approach*, Boca Raton, FL: Chapman & Hall/CRC, 2nd ed.

Gill, J. and Walker, L. D. (2005), "Elicited Priors for Bayesian Model Specifications in Political Science Research," *The Journal of Politics*, 67, 841–872.

Herrnson, P. S. and Morris, I. L. (2007), "Presidential Campaigning in the 2002 Congressional Elections," *Legislative Studies Quarterly*, 32, 629–648.

Holland, P. W. (1986a), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

— (1986b), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Imai, K. and Yamamoto, T. (2008), "Causal Inference with Measurement Error: Nonparametric Identification and Sensitivity Analyses of a Field Experiment on Democratic Deliberations," Paper presented at the 2008 Summer Political Methodology Meeting.

Jackman, S. (2000), "Estimation and Inference via Bayesian Simulation: an Introduction to Markov Chain Monte Carlo," *American Journal of Political Science*, 44, 375–404.

Kadane, J. B. and Wolfson, L. J. (1998), "Experiences in Elicitation," *The Statistician*, 47, 3–19.

Keele, L. J., Fogarty, B., and Stimson, J. A. (2004), "The Impact of Presidential Visits in the 2002 Congressional Elections," *PS: Political Science & Politics*, 34, 971–986.

King, G. (1997), *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton: Princeton University Press.

King, G., Tomz, M., and Wittenberg, J. (2000), "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science*, 44, 347–361.

Manski, C. F. (1990), "Nonparametric Bounds on Treatment Effects," *The American Economic Review Papers and Proceedings*, 80, 319–323.

— (1993), "Identification Problems in the Social Sciences," *Sociological Methodology*, 23, 1–56.

— (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.

— (1997), "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

— (2003), *Partial Identification of Probability Distributions*, New York: Springer.

Neustadt, R. E. (1960), *Presidential Power*, New York, NY: Wiley.

Nutting, B. and Stern, A. H. (2001), *CQ's Politics in America: 2002 the 107th Congress*, Washington, D.C.: Congressional Quarterly Press.

Pearl, J. (1995), "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–710.

— (2000), *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press.

Richardson, T. (2004), "Discussion on the Paper by Wakefield," *Journal of the Royal Statistical Society: Series A*, 167, 438.

Robins, J. M. (1986), "A new aproach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect," *Mathematical Modeling*, 7, 1393–1512.

— (1989), "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. Sechreset, L., Freeman, H., and Mulley, A., Washington DC: U.S. Public Health Service.

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

— (1990), "Formal Modes of Statistical Inference For Causal Effects," *Journal of Statistical Planning*, 25, 279–292.

Sellers, P. J. and Denton, L. M. (2006), "Presidential Visits and Midterm Senate Elections," *Presidential Studies Quarterly*, 36, 410–432.

Western, B. and Jackman, S. (1994), "Bayesian Inference for Comparative Research," *American Political Science Review*, 88, 412–423.

Appendices

A Prior Selection and Patterns of Confounding

Here we fully elaborate how various prior values govern different patterns of potential outcomes under general confounding. Table 6 provides an overview of the various parameters in the model. It is by placing prior distributions on the $\boldsymbol{\psi}$ parameters that we use subjective knowledge to elicit possible causal patterns.

Parameter θ_{xy}	Probability $\Pr(X_i = x, Y_i = y)$	Interpretation Probability X_i is equal to x and Y_i is equal to y
ψ_{00}	$\Pr(Z_i = 1 \mid X_i = 0, Y_i = 0)$	Probability i would be helped by treatment given i not treated and i failed
$1 - \psi_{00}$	$\Pr(Z_i = 0 \mid X_i = 0, Y_i = 0)$	Probability i would never succeed given i not treated and i failed
ψ_{01}	$\Pr(Z_i = 3 \mid X_i = 0, Y_i = 1)$	Probability i would always succeed given i not treated and i succeeded
$1 - \psi_{01}$	$\Pr(Z_i = 2 \mid X_i = 0, Y_i = 1)$	Probability i would be hurt by treatment given i not treated and i succeeded
ψ_{10}	$\Pr(Z_i = 2 \mid X_i = 1, Y_i = 0)$	Probability i was hurt by treatment given i treated and i failed
$1 - \psi_{10}$	$\Pr(Z_i = 0 \mid X_i = 1, Y_i = 0)$	Probability i would never succeed given i treated and i failed
ψ_{11}	$\Pr(Z_i = 3 \mid X_i = 1, Y_i = 1)$	Probability i would always succeed given i treated and i succeeded
$1 - \psi_{11}$	$\Pr(Z_i = 1 \mid X_i = 1, Y_i = 1)$	Probability i was helped by treatment given i treated and i succeeded

Table 6: *Interpretation of Parameters in the Model for (X,Y,Z).* The i indices denote a randomly selected unit.

Begin with the situation in which $X_i = 0$ and $Y_i = 0$. Here two potential outcome profiles are possible: $Z_i = 0$ (never succeed) and $Z_i = 1$ (helped). ψ_{00} is the conditional probability that $Z_i = 1$ (i would be helped by treatment) given $X_i = 0$ and $Y_i = 0$, while $1 - \psi_{00}$ is obviously the conditional probability that $Z_i = 0$ (i would never succeed) given $X_i = 0$ and $Y_i = 0$. b_{00} is the number of pseudo $Z = 1$ (helped) observations + 1 and c_{00} is the

number of pseudo $Z = 0$ (never succeed) observations + 1. If our background knowledge suggests that units for which we observe $X = 0$ and $Y = 0$ are unlikely to respond to treatment we would set $c_{00} > b_{00}$. This implies that $\Pr(Z_i = \text{ never succeed } | X_i = 0, Y_i = 0) > \Pr(Z_i = \text{ helped } | X_i = 0, Y_i = 0)$. On the other hand, if we believe that these units are more likely than not to respond to treatment we would set $b_{00} > c_{00}$. This would imply $\Pr(Z_i = \text{ helped } | X_i = 0, Y_i = 0) > \Pr(Z_i = \text{ never succeed } | X_i = 0, Y_i = 0)$. The absolute magnitude of b_{00} and c_{00} determines how sure we are of the potential outcome distribution within the $X = 0$, $Y = 0$ group.

Next consider the situation in which $X_i = 0$ and $Y_i = 1$. Again, two potential outcome profiles are possible: $Z_i = 2$ (hurt) and $Z_i = 3$ (always succeed). ψ_{01} is the conditional probability that $Z_i = 3$ (i would always succeed) given $X_i = 0$ and $Y_i = 1$, while $1 - \psi_{01}$ is the conditional probability that $Z_i = 2$ (i would be hurt by treatment) given $X_i = 0$ and $Y_i = 1$. b_{01} is the number of pseudo $Z = 3$ (always succeed) observations + 1 and c_{01} is the number of pseudo $Z = 2$ (hurt) observations + 1. If our background knowledge suggests that units for which we observe $X = 0$ and $Y = 1$ are unlikely to respond to treatment we would set $b_{01} > c_{01}$. On the other hand, if we believe that these units are more likely than not to respond (negatively) to treatment we would set $c_{01} > b_{01}$. Again, the absolute magnitude of b_{01} and c_{01} determines how sure we are of the potential outcome distribution within the $X = 0$, $Y = 1$ group.

Within the $X_i = 1$ and $Y_i = 0$ group the two potential outcome profiles are $Z_i = 0$ (never succeed) and $Z_i = 2$ (hurt) . ψ_{10} is the conditional probability of randomly selecting a unit for which $Z_i = 2$ (i is hurt by treatment) from the $X_i = 1$ and $Y_i = 0$ group. b_{10} is the number of pseudo $Z = 2$ (hurt) observations + 1 and c_{10} is the number of pseudo $Z = 0$ (never succeed) observations + 1. Setting $b_{10} > c_{10}$ would be consistent with a prior belief that more subjects tend to respond negatively to treatment within this group, while setting $c_{10} > b_{10}$ would be consistent with a belief that treatment is more likely than not to have no effect on units within this group.

Finally, within the $X_i = 1$ and $Y_i = 1$ group we see that the two potential outcome profiles are $Z_i = 1$ (helped) and $Z_i = 3$ (always succeed). ψ_{11} is the conditional probability of seeing a $Z_i = 3$ (always succeed) observation within this group. b_{11} is the number of pseudo $Z = 3$ (always succeed) observations + 1 and c_{11} is the number of pseudo $Z = 1$ (helped) observations + 1. If one thinks that units within this group are, on average, likely to respond positively to treatment one would set $c_{11} > b_{11}$. If non-responsiveness is hypothesized one would set $b_{11} > c_{11}$.

B Prior Selection in the Application

The treatment effect estimates reported required specification of the prior distributions for the ψ parameters which govern the mixing fractions due to an unobserved confounder. Table 7 contains the values used for the priors associated with the various assumptions used in the analysis. One key question, of course, is how did we select these particular values and would the conclusions drawn in the analysis change if different values were used? For the monotonicity assumption, this is relatively easy. First, a stronger monotonicity assumption is not possible since the difference in the beta distribution parameters for ψ_{01} and $\psi 10$ cannot be any greater than those used. The sampling process for the posterior estimates will not converge with a larger spread on the beta distributions. At the other end for monotonicity to hold, there inequalities outlined in the paper must hold. A difference of five points in the beta distribution parameters is the smallest spread where the bounds include zero. We then selected prior values roughly half way between these two prior specifications.

For the selection priors, the estimates were largely invariant to the choice of prior parameters. We chose an upper bound on the effect size based on which selection prior combined with the strong monotonicity assumption produced and estimate equal to the naive treatment effect. We then chose evenly spaced increments for the prior parameters that preserved the inequality required for the selection assumption to hold. While larger spreads on these parameters could be used, it is unrealistic to use a combination of assumptions that pro-

duces treatment effect estimates that are larger than the naive estimate. Moreover, for monotonicity there is a clear upper-bound to the identifying power of that assumption.

Table 7: Prior Values Used in Analysis

	Monotonicity	
	$b_{01} > c_{01}$	$b_{10} < c_{10}$
Weak	10,5	5,10
Medium	10,1	10,1
Strong	10, 0.02	0.02, 10
	Selection	
	$b_{00} > c_{00}$	$b_{11} < c_{11}$
Weak	10, 8	8, 10
Medium	11, 7	7, 11
Strong	13, 6	13, 6