Bias Amplification and Bias Unmasking

Joel A. Middleton[1], Marc A. Scott[2], Ronli Diakow[2] and Jennifer L. Hill[2]

[1]Department of Political Science, University of California, Berkeley
[2]Humanities and Social Sciences in the Professions, New York University, Steinhardt

June, 2015

ABSTRACT

In the analysis of causal effects in non-experimental studies, conditioning on observable covariates is one way to try to reduce unobserved confounder bias. However, a developing literature has shown that conditioning on certain covariates may increase bias, but the mechanisms underlying this phenomenon have not been fully explored. We contribute to the literature on bias-increasing covariates by first introducing a way to decompose omitted variable bias into constituent parts: bias due to an unobserved confounder, bias due to *excluding* observed covariates, and bias due to amplification. This leads to two important findings. First, we identify the fact that the popular approach of adding fixed-effects can lead to bias amplification, even while they are not instruments. Instruments have been the focus of the bias amplification literature to date. That fixed effects might amplify (or otherwise increase) bias may be unexpected because fixed effects are often thought to be a convenient way to account for any and all group-level confounding. Second, we introduce the concept of bias unmasking and show how it can be even more insidious than bias amplification in some cases. After introducing these new results analytically, we use constructed observational placebo studies to illustrate bias amplification and bias unmasking with real data. Finally, we propose a way to add bias decomposition information to sensitivity analysis graphical displays to help practitioners think through the potential for bias amplification and bias unmasking in actual applications.

This research was partially supported by Institute of Education Sciences grants R305D110037 and R305B120017.

1. INTRODUCTION

In the analysis of causal effects in non-experimental studies the key assumption necessary for unbiased estimation is that all confounders (pre-treatment variables that predict both treatment assignment and the outcome) have been measured. In the social science literature this assumption is often referred to as “selection on observables” (Heckman and Robb 1985, 1986), “conditional independence” (Lechner 2001) or “ignorability” (Rubin 1978), and it is well-known that violation of this assumption leads to biased inference. However it is typically implausible to believe that we have measured all confounders, raising the question as to which of the available covariates should be adjusted for (conditioned upon) in practice.

Advice in the extant literature on which variables to condition on are contradictory. One recommendation has been that conditioning on more, rather than fewer, available (pretreatment) covariates is the best way to minimize bias associated with unobserved sources of heterogeneity (Rubin 2002; Rosenbaum 2002). Another recommendation says that those variables that are related to the treatment assignment mechanism should be included in the analysis (D’Agostino, Jr. 1998). Still other advice is to choose covariates based on their relationship to the outcome, rather than to the treatment (Brookhart et al. 2010; Austin et al. 2007; Hill 2007a).

There are, however, two notable classes of covariates that most agree should be *excluded* from the set of conditioning covariates. These are *bias inducers* and *bias amplifiers*. Bias inducers include posttreatment variables such as mediators and colliders (Cole et al. 2010; Pearl 2000; Schisterman et al. 2009), and a particular group of pretreatment covariates (pretreatment colliders that lead to M-bias or butterfly-bias) (Ding and Miratrix 2014; Sjlander 2009; Pearl 2009). Such bias inducers may not be troublesome in practice, however, either because they can be identified for exclusion, as is sometimes the case for posttreatment variables, or because the bias they induce tends to be small (Ding and Miratrix 2014; Liu et al. 2012; Greenland 2002).

Bias amplifiers have received recent attention as variables that should be excluded from the conditioning set (Pearl 2010; Wooldridge 2009; Bhattacharya and Vogt 2007; Pearl 2011; Myers et al. 2011; Wyss et al 2014). These covariates cannot induce bias where there is none, but they increase bias by modifying bias that is due to an unobserved confounder. Instruments, variables related to treatment but not directly causing the outcome, are the canonical example of a bias amplifying covariate. Conditioning on an instrument can hurt but can never help. On the one hand, this may seem like a trivial concern because it is unclear under what circumstances a researcher would be unaware that a variable was a true instrument for their treatment variable. However, even imperfect instruments can amplify bias (cf. Pearl 2010) and, as we will show below, even noninstruments can amplify bias.

In sections 2 and 3 we show how omitted variable bias may be decomposed into several constituent parts: bias attributable to the unobserved confounder, bias due to omitting observed covariates, and bias due to amplification. Doing so allows us to make two contributions regarding bias increasing covariates. The first, in section 4, we show that fixed effects can act as pure bias amplifiers. Fixed effects are not instruments, which have been the focus of the bias amplification literature to date. Moreover, fixed effects are often thought to be useful for absorbing unmeasured group-level confounding, so demonstrating that they can increase bias in general may be unexpected. The second contribution is the introduction of the conept of bias "unmasking", which helps to frame why even variables that do not amplify bias per se may still lead to net increased bias. In sections 5 and 7, we examine two case studies where the causal effect is known and where confounding is likely to be present to estimate and decompose biases into the constituent parts. In one case study amplification is a major contributor to net bias. In the other, the inclusion of covariates leads to a larger *net* bias due to "unmasking" of unobserved confounder bias. These examples reinforce prior advice to avoid inadvertently controlling for instruments when trying to infer causal effects from data where the causal variable was not randomized. They also illustrate why applied researchers might be concerned about arbitrary use of fixed effects in non-experimental studies. In section **??**, we provide a suggested modification to sensitivity analysis graphical displays to help practitioners think about the potential for amplification and unmasking.

Overall, we argue that none of the extant recommendations for the practice of identifying a conditioning set of covariates provide much help to researchers in light of our analytic and empirical analyses. In particular, we demonstrate that the common advice that inclusion of fixed effects is a low-risk strategy to reduce bias may be misguided[1]. We provide some assistance by describing the potential role of sensitivity analysis in illuminating which variables might act as bias amplifiers or bias unmaskers. This approach does require some prior information about the nature of the unobserved confounder, suggesting that researchers must bring substantive knowledge to the table when determining whether a covariate should be included.

2. WHETHER TO CONDITION ON X

In this section we will establish the conditions under which a researcher would want to condition on a set of covariates, X, in estimating the effect of a treatment, Z, on an outcome, Y.

Mathematically describing the magnitude of bias incurred by failure to satisfy the selection on observables assumption requires additional assumptions about the relationships between variables. We derive our results using the linear model as has been done in related work (Ding and Miratrix 2014; Pearl 2010; Clarke 2005, 2009) and which has the advantage of tying this work into more general results regarding omitted variable bias.[2]

To proceed, consider a linear model relating an outcome, Y, and a treatment, Z,

$$Y = Z\tau + X\beta^y + U\zeta^y + \epsilon^y. \tag{1}$$

[1] For a distinct problem with use of fixed effects for causal inference see Sobel (2006).

[2] While we focus on the linear model, we expect that many of the results hold, broadly speaking, for GLM models. However, GLM models come with a host of problems of their own with respect to bias. For example, coefficients from unadjusted and covariate adjusted logistic regression models are not comparable (VanderWeele and Arahc 2011; Freedman 2008; Breen et. al. 2013), a problem sometimes referred to as "noncollapsibility" (cf. VanderWeele 2015). Therefore, a lengthy discussion of the bias of adjusted and unadjusted GLM estimators is beyond the scope of this paper.

Here X is a matrix of observed covariates, and U is an unobserved confounder. As in Carnegie et al. (2014a) and Imbens (2003) we make the simplifying assumption that $U \perp X$. We can justify this assumption by conceptualizing U as the portion of the unobserved covariate that is orthogonal to the observed covariates. For the sake of clarity, and without loss of generality, we also assume that $\mathrm{E}[Z] = 0$.

Next we derive bias both for the case where X is omitted from the conditioning set and when it is included in the conditioning set so that the two biases might be compared and show how to decompose the biases into constituent parts to facilitate the comparison. To do so, we refer to Appendix A where we derive a (well known) generic expression for omitted variable bias.

First suppose one calculates the unadjusted estimate of τ in equation (1) by simply regressing Y on Z. Substituting $S = [Z]$ and $O = [X\ U]$ in (14) in Appendix A yields an expression for the omitted variable bias of the crude estimator,

$$\begin{aligned} \text{Bias}\left[\widehat{\tau}_{\text{Y|Z}}\right] =& \mathrm{E}\left[\left(Z'Z\right)^{-1} Z'X\beta^y + \left(Z'Z\right)^{-1} Z'U\zeta^y\right] \\ =& \chi + \upsilon \end{aligned} \tag{2}$$

where $\chi \equiv \mathrm{E}\left[\left(Z'Z\right)^{-1} Z'X\beta^y\right]$ is the bias due to omitting X and $\upsilon \equiv \mathrm{E}\left[\left(Z'Z\right)^{-1} Z'U\zeta^y\right]$ is the bias due to omitting U. The symbols χ and υ are used as a shorthand to signify the constituent parts of the bias.

Now suppose we are interested in the bias when estimating τ in a new model that includes X in the conditioning set, S. In that case, substitute $S = [Z\ X]$ and $O = [U]$ in (14) in Appendix A. The bias can be written in partitioned notation as follows:

$$\text{Bias}\begin{bmatrix} \widehat{\tau}_{\text{Y|ZX}} \\ \widehat{\beta}^y \end{bmatrix} = \mathrm{E}\left[\begin{bmatrix} Z'Z & Z'X \\ X'Z & X'X \end{bmatrix}^{-1} \begin{bmatrix} Z' \\ X' \end{bmatrix} U\zeta^y\right]. \tag{3}$$

Using the inverse of the partition matrix and selecting off the element that corresponds to the coefficient on the causal variable Z (steps in Appendix A), write

$$\begin{aligned}
\text{Bias}\left[\widehat{\tau}_{\text{Y|ZX}}\right] &= \text{E}\left[\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^{y}\right] \\
&\doteq \left(\frac{1}{1-r^2_{Z|X}}\right)\upsilon \\
&= \upsilon + \left(\frac{r^2_{Z|X}}{1-r^2_{Z|X}}\right)\upsilon \\
&= \upsilon + \alpha
\end{aligned} \tag{4}$$

where $r^2_{Z|X}$ is the coefficient of determination, R-squared, in the regression of Z on X and $\alpha \equiv \left(\frac{r^2_{Z|X}}{1-r^2_{Z|X}}\right)\upsilon$. Note in the second line of (4) the dot above the equal sign, signifying an approximate equality. The approximation follows from the fact that the ratio $\left(\frac{1}{1-r^2_{Z|X}}\right)$ has been placed in front of the expectation operator in the term $\upsilon = \text{E}\left[(X'X)^{-1}X'U\zeta^{y}\right]$. The term $\left(\frac{1}{1-r^2_{Z|X}}\right)$ can be referred to as the amplification factor; importantly this term is identified and thus can be estimated by the researcher. The amplification factor is particularly problematic if X accounts for a great deal of variation in Z as noted by Pearl (2010). The term $\alpha \equiv \left(\frac{r^2_{Z|X}}{1-r^2_{Z|X}}\right)\upsilon$ gives the change in bias attributable to amplification, call it the *net amplification bias*.

A careful comparison of (4) and (2), reveals two key insights about adding X to the conditioning set of covariates. First, note the bias term in (2) associated with omitting X, namely, $\chi \equiv \text{E}\left[(Z'Z)^{-1}Z'X\beta^{y}\right]$. This term is absent in (4) because X is adjusted for in this model.

Second, the bias due to omitting U is modified from $\text{E}\left[(Z'Z)^{-1}Z'U\zeta^{y}\right]$ in (2) to become $\text{E}\left[\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^{y}\right]$ in (4). The difference between these two terms results in the appearance of $-Z'X\left[X'X\right]^{-1}X'Z$ in the denominator, a term which is necessarily less than or equal to zero because it is (-1 times) a quadratic form with positive definite matrix $\left[X'X\right]^{-1}$ (cf. Greene 2000, sections 2.8 and 2.8.1). Therefore, except if the term $-Z'X\left[X'X\right]^{-1}X'Z$ is zero (i.e., X is not correlated with Z), it shrinks the denominator in (4) relative to (2) resulting in *amplification* of the bias due to the unobserved U.

It is useful at this juncture to consider why instrumental variables (Angrist et al. 1996) have been a particular focus of attention in discussing bias amplifiers (Pearl 2010; Wooldridge 2009; Bhattacharya and Vogt 2007). When X is an instrument, $\beta^y = 0$ by definition. Therefore the bias due to omitting X, χ, in (2) equals 0 and there can be no benefit due to removing χ bias when going from (2) to (4) and the only change in bias is an increase due to amplification, α. In that sense, instruments can be referred to as *pure* amplifiers. When X is a pure amplifier it is necessarily true that (4) is larger than (2).

However, amplification is only part of the story about the change in bias when going from an unadjusted to adjusted estimator. Whether conditioning on X increases or decreases the *net* bias depends on the magnitude of (2) relative to the magnitude of (4). Formally, a set of covariates, X, can be said to be net bias reducing only when

$$\left|\mathrm{E}\left[(Z'Z)^{-1}Z'U\zeta^y + (Z'Z)^{-1}Z'X\beta^y\right]\right| > \left|\mathrm{E}\left[\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^y\right]\right|.$$

Or write,

$$|\upsilon + \chi| > |\upsilon + \alpha|\,.$$

If υ (the bias due to omitting U) and χ (the bias due to omitting X) have the same sign then this implies that for X to be bias reducing

$$|\chi| > |\alpha|\,.$$

When υ and χ have opposite signs the requirement for X to be bias reducing is

$$|\chi| > |2\upsilon + \alpha| \, .$$

Clearly, conditioning on X can be net bias increasing in cases where the bias due to amplification, α, is relatively large. However, conditioning on X can be net bias increasing even when $r^2_{Z|X} = 0$ (and, hence, α=0) if the bias due to omitting U, υ, and the bias due to omitting X, χ, have opposite sign and $|\chi| < 2\,|\upsilon|$. In that case, because χ has an opposite sign to υ but similar magnitude, it can be said to be *masking* (or canceling) υ in (2). In that case the χ is a "good" bias because it cancels with υ, rendering the net bias of the unadjusted estimator closer to zero than that of the adjusted estimator.

That bias due to omitting a known covariate X can be "good" bias (because it masks bias due to the unobserved confounder) is troubling because it implies that even when X is known to be predictive of Y, including it in the conditioning set of covariates may increase overall bias. To know whether removing χ bias improves net bias or not, one must know something about υ which is not identified. In light of this observation, it is clear that none of the existing recommendations for practice provide complete guidance on whether to condition on a covariate, or set of covariates (as with fixed effects), or not.

3. WHETHER TO CONDITION ON X_1 GIVEN THAT X_2 WILL BE INCLUDED IN THE CONDITIONING SET

In this section we generalize the above results to the case where we want to know whether to include *all* of the covariates in X in the conditioning set given that *some* of them will be included in the conditioning set. Notationally, first partition the matrix of covariates such that $X = [X_1 X_2]$. Now assume that X_2 will certainly be in the conditioning set and the question is whether to also include X_1 in the conditioning set. As we show, the results for bias amplification are analogous to the simplified case above.

The model can now be written,

$$Y = Z\tau + X_1\beta_1^y + X_2\beta_2^y + U\zeta^y + \epsilon^y \tag{5}$$

where the U is independent of both X_1 and X_2. Omitting X_1 from the conditioning set leads to the bias

$$\begin{aligned}
\text{Bias}\left[\widehat{\tau}_{\text{Y}|\text{ZX}_2}\right] &\doteq \left(\frac{1}{1-r^2_{Z|X_2}}\right) \text{E}\left[\left(Z'Z\right)^{-1} Z'X_1\beta_1^y + \left(Z'Z\right)^{-1} Z'U\zeta^y\right] \\
&= \left(\frac{1}{1-r^2_{Z|X_2}}\right)\left[\chi_1 + \upsilon\right] \\
&= \chi^* + \upsilon^*
\end{aligned} \tag{6}$$

where $\upsilon^* \equiv \left(\frac{1}{1-r^2_{Z|X_2}}\right)\upsilon$, likewise $\chi^* \equiv \left(\frac{1}{1-r^2_{Z|X_2}}\right)\chi_1$ and $r^2_{Z|X_2}$ is the R-squared in the regression of Z on X_2. As before, the dot above the equal sign in the first line of equation (6) is due to the fact that $\left(\frac{1}{1-r^2_{Z|X_2}}\right)$ is outside the expectation operator.

Including X_1 in the conditioning set leads to the bias

$$\begin{aligned}
\text{Bias}\left[\widehat{\tau}_{\text{Y}|\text{ZX}_1\text{X}_2}\right] &\doteq \left(\frac{1}{1-r^2_{Z|X_1X_2}}\right)\upsilon^* \\
&= \upsilon^* + \left(\frac{r^2_{Z|X_1X_2} - r^2_{Z|X_2}}{1-r^2_{Z|X_1X_2}}\right)\upsilon^* \\
&= \upsilon^* + \alpha^*
\end{aligned} \tag{7}$$

where $r^2_{Z|X_1X_2}$ is the R-squared in the regression of Z on X_1 and X_2. Here the net amplification bias, $\alpha^* \equiv \left(\frac{r^2_{Z|X_1X_2} - r^2_{Z|X_2}}{1-r^2_{Z|X_1X_2}}\right)\upsilon^*$, is defined only slightly differently from α above and the amplification factor can be written $\left(\frac{1-r^2_{Z|X_2}}{1-r^2_{Z|X_1X_2}}\right)$.

So, conditioning on X is bias reducing when

$$|\upsilon^* + \chi^*| > |\upsilon^* + \alpha^*|\,.$$

When υ^* and χ^* have the same sign the requirement is that $|\chi^*| > |\alpha^*|$. When υ^* and χ^* have different signs the requirement is that $|\chi^*| > |2\upsilon^* + \alpha^*|$.

These results are analogous to the simpler case of a single block of covariates, X, considered in the section above.

4. FIXED EFFECTS AS PURE BIAS AMPLIFIERS

As mentioned above, pure bias amplifiers such as instruments can be particularly problematic because there cannot be any benefit to removing χ from the bias equation since $\chi \equiv 0$. In this section we identify the conditions where fixed effects can be pure amplifiers, amplifying bias but providing no net improvement in bias due to removing χ bias.

To consider fixed effects under the rubric presented above simply imagine X as a matrix of dummy variables. Starting from this point of view, the term $\chi \equiv \mathrm{E}\left[(Z'Z)^{-1} Z'X\beta^y\right]$ in (2) can be written $\chi \equiv \mathrm{E}\left[\sum_{k=1}^{K} (Z'Z)^{-1} Z'X_k\beta^{yk}\right]$ where X_k is the column vector from X associated with the k^{th} dummy variable.

Now consider the case where fixed effects are pure amplifiers – when the term
$\chi \equiv \mathrm{E}\left[\sum_{k=1}^{k} (Z'Z)^{-1} Z'X_k\beta^{yk}\right] \equiv 0$. Trivially, this term can be zero if the terms β^{yk} are all zero, i.e., if the fixed effects are instruments, but it can also be zero because the positive and negative terms sum to zero.

When might those positive and negative terms net out to zero? To develop an intuition, consider a model for the treatment, Z,

$$Z = \sum_{k=1}^{K} X_k \beta^{zk} + U\zeta^z + \epsilon^z \tag{8}$$

where U is defined as above. Now make the assumption, for the sake of simplifying the exposition, that Z has unit variance (in addition to having mean zero) and that the size of the K groups (associated with the fixed effects) are equal and thus $E[X_k] \equiv 1/K$. Then we might write for the i^{th} dummy variable

$$\begin{aligned}
\left(Z'Z\right)^{-1} Z'X_k &= \text{Cov}(X_k, Z) \\
&= \text{Cov}(X_k, \sum_j X_j\beta^{zj} + U\zeta^z + \epsilon^z) \\
&= \frac{1}{K}\beta^{zk} - \frac{1}{K^2}\sum_{j=1}^{K}\beta^{zj} \qquad (9)
\end{aligned}$$

So,

$$\begin{aligned}
\chi \equiv \sum_{k=1}^{K} \left(Z'Z\right)^{-1} Z'X_k\beta^{yk} &= \sum_{k=1}^{K} \text{Cov}(X_k, Z)\beta^{yk} \\
&= \sum_{k=1}^{K}\left(\frac{1}{K}\beta^{zk} - \frac{1}{K^2}\sum_{j=1}^{K}\beta^{zj}\right)\beta^{yk} \\
&= \frac{1}{K}\sum_{k=1}^{K}\beta^{zk}\beta^{yk} - \frac{1}{K^2}\sum_{k=1}^{K}\left(\sum_{j=1}^{K}\beta^{zj}\right)\beta^{yk} \\
&= E_k\beta^{zk}\beta^{yk} - E_k\beta^{zj}E_k\beta^{yk} \\
&= Cov_k\left(\beta^{zk}, \beta^{yk}\right) \qquad (10)
\end{aligned}$$

where use the notation Cov_k to denote that covariance is to be taken *across* the K groups. Likewise, E_k is expectation across the K groups.

The derivation shows that fixed effects can be pure bias amplifiers when $Cov_k\left(\beta^{zk}, \beta^{yk}\right) = 0$.

One way to interpret this condition is that the group level structure in Y does not covary with the group level structure in Z.

At first look, having ascertained this result would seem to provide some encouragement. If the expression in the last line of (10) can be estimated, then the fixed effects can be avoided when they are pure bias amplifiers (or close to it). Sadly, the term cannot be estimated unbiasedly, or even meaningfully bounded. We can see this by examining Bias$[\widehat{\beta^y}]$ in (17) in the Appendix A. Instead, the usual regression estimator $\widehat{\beta^{yk}}$ converges in probability to

$$\beta^{yk} - \beta^{zk}\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^y. \tag{11}$$

Asymptotically then, a quantity that one might estimate is

$$\begin{aligned}
\widehat{Cov}_k\left(\beta^{zk},\beta^{yk}\right) &= Cov_k\left(\beta^{zk},\widehat{\beta}^{yk}\right)\\
&= Cov_k\left(\beta^{zk},\beta^{yk} - \beta^{zk}\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^y\right)\\
&= Cov_k\left(\beta^{zk},\beta^{yk}\right) - V_k\left(\beta^{zk}\right)\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^y\\
&= Cov_k\left(\beta^{zk},\beta^{yk}\right) - V_k\left(\beta^{zk}\right)(\upsilon+\alpha)
\end{aligned} \tag{12}$$

where $V_k(\beta^{zk})$ represents the variance of the values of β^{zk}. Because $(\upsilon+\alpha)$ may take on a potentially wide range of values (12) is not a useful estimator of (10).[3]

5. CASE STUDY - THE EFFECT OF A GET-OUT-THE-VOTE INTERVENTION

In this section we repurpose the data from a study of the effect of prerecorded get-out-the-vote phone calls on voter turnout (Shaw et al. 2012) to illustrate the phenomenon of bias amplification.

[3] That said, in a sensitivity analysis framework then, estimates for $Cov_k\left(\beta^{zk},\beta^{yk}\right)$ might be computed for posited values of $\left(Z'Z - Z'X\left[X'X\right]^{-1}X'Z\right)^{-1}Z'U\zeta^y$ – the bias term in (4).

While the original study was a randomized experiment, we use the data to create a constructed observational study.

In the original experiment units were assigned to a condition that received a prerecorded telephone message encouraging them to vote or to a “no message” condition. In 1,597 precincts randomization was at the precinct level. In another 5,838 precincts, households were randomly assigned to treatment or control within precinct. We utilize the combined data file of 463,489 subjects.

Of interest in this study was the effect of contact on voter turnout. However, individuals who were actually contacted may be different from those assigned to treatment who were not contacted in ways that make them more likely to vote – for instance they were less likely to have died or moved. Therefore, naively regressing turnout on contact is likely to violate the selection on observables assumption and thus yield a biased estimate of the effect of contact. Instrumental variables regression, using treatment assignment as the instrument, is the typical remedy. However, we are interested in illustrating bias so we do not use instrumental variables. Instead we deliberately produce biased estimates. Moreover, we construct a placebo test, using turnout in *prior* elections as outcome measures. Since we know that contact in 2006 cannot affect the turnout in prior elections, the true treatment effect must be zero. Estimates that deviate from zero thus reveal the bias inherent in the estimator.[4]

Within the context of our constructed observational placebo study we can test whether two types of variables act as bias amplifiers when included as covariates in the specified model. In section 5.1 we consider treatment assignment (an instrument for contact) as a bias amplifier. In section 5.2 we consider fixed effects for precinct as bias amplifiers. In both sections, estimates from models that include the potential bias amplifier are compared to a simple regression of turnout on

[4]Another option would have been to use 2006 election turnout as the outcome and compare our observational estimates to the experimental benchmark created by the instrumental variables estimate. The downside of this approach is that this benchmark is itself noisy making it more difficult to precisely partition the bias. We prefer using the sharp 0 of our placebos tests as a comparison.

contact to see which yields an estimate closer to the true parameter value of zero. If a model with the potential bias amplifier yields an estimate that is further from zero, then this is evidence that the potential bias amplifier caused a net increase in bias. Furthermore, because the causal parameter is known to be zero, the constituent components of bias – χ, υ and α – are also identified. So for each outcome we can see if the bias amplification is the cause of the net increase in bias.

5.1. *Instrument as Bias Amplifier: Analysis and Results*

Table 5.1, panel A shows the results for the analysis of the effect of the instrument, randomized treatment assignment, on bias.

To describe the model specification we refer back to model (1). Y is an $(n \times 1)$ vector of voter turnout indicators, Z is an $(n \times 1)$ vector of indicators for contact, X is an $(n \times 1)$ vector of indicators of treatment assignment. U is the omitted confounder, assumed to have unit variance. Each row of the table conducts the analysis for a different election. The column labeled 'OLS' presents the estimated coefficient on Z when regressing Y on Z only. The column labeled 'Inst.' presents the estimated coefficient on Z when regressing Y on X and Z. The column 'Diff' presents the difference between the two estimates along with a bootstrapped standard error. In the columns labeled υ, α, and χ the observed bias is decomposed into constituent parts.

For the general election 2004, the OLS estimate exhibits a bias of 0.138 while the model controlling for treatment is much more biased at 0.478.[5],[6] The bias increase of 0.339 (an increase of 244%) is entirely due to bias amplification, α. That there is essentially no contribution to the bias through χ is expected given that instruments are known to be pure amplifiers. Not surprisingly then, the unadjusted estimator is better than one that adjusts for an instrument.

Results in Table 5.1, Panel A, from other election years show substantively similar results. As

[5]The standard errors are so small as to suggest that the bias is measured with great precision.
[6]This is a tremendous amount of bias when one considers that the outcome is a binary, 0-1, outcome.

in the case of the General Election 2004, adding the treatment indicator to the conditioning set of covariates leads to increased bias. The increase in bias is attributable to bias amplification.

Results in Table 5.1, Panel B, repeat this analysis for models that include additional covariates (turnout in prior elections) in the conditioning set. To describe this model specification we refer back to model (5). Y and Z are defined as in Panel A. X_1 is the instrument while X_2 is an $(n \times k)$ matrix of indicators for turnout in k prior elections. For the General 2004 election outcome, X_2 included General 2002 turnout, General 2000 turnout, Primary 2004 turnout, Primary 2002 turnout and Primary 2000 turnout. For the Primary 2004 election outcome, X_2 included General 2002 turnout, General 2000 turnout, Primary 2002 turnout and Primary 2000 turnout. The column labeled OLS presents the estimated coefficient on Z when regressing Y on Z and X_2. In the next column, labeled Inst., is the estimated coefficient on Z when regressing Y on Z, X_2 and also X_1. The remaining columns give the difference between the two estimates and the bias decomposition.

Overall the biases are smaller in Panel B. For example, in Panel A, for General 2004 the bias due to omitting U, υ, is estimated to be 0.140, or 14 percentage points. In Panel B, in contrast, the bias due to omitting U, υ^* is estimated to be 0.021 or two percentage points. However, we note that 2.1 percentage points is still a substantively large bias. Moreover, the ratio $\frac{\alpha^*}{\upsilon^*}$ is similar to $\frac{\alpha}{\upsilon}$ above; bias due to amplification, α^*, is over 250% the size of the bias due to omitting U, υ^*.

5.2. *Fixed Effects as Bias Amplifiers: Analysis and Results*

Next, consider the implications for bias when adding fixed effects for precinct to the model specification. Table 2 presents these results. In Panel A, referring back to model (1), Y and Z are defined as the turnout indicators and contact indicators, as above, while X is now an $(n \times K)$ matrix of dummy variable indicators for the K precincts.

Examining the results for 2004 election turnout, the fixed effects model is much more biased than the model without fixed effects; when regressing Y on Z only the estimate is 0.138 com-

A. No covariates	OLS (SE)	Inst. (SE)	Diff (SE)	υ	α	χ
General 2004	0.138	0.478	0.339	0.140	0.337	-0.002
	(0.004)	(0.004)	(0.003)			
General 2002	0.135	0.451	0.315	0.132	0.318	0.003
	(0.005)	(0.004)	(0.004)			
General 2000	0.131	0.451	0.32	0.132	0.318	-0.001
	(0.005)	(0.004)	(0.004)			
Primary 2004	0.094	0.285	0.192	0.084	0.202	0.01
	(0.005)	(0.004)	(0.004)			
Primary 2002	0.093	0.288	0.195	0.085	0.204	0.009
	(0.004)	(0.003)	(0.004)			
Primary 2000	0.113	0.37	0.257	0.109	0.261	0.004
	(0.004)	(0.004)	(0.003)			
B. With covariates	**OLS (SE)**	**Inst. (SE)**	**Diff (SE)**	υ^*	α^*	χ^*
General 2004	0.017	0.077	0.06	0.021	0.056	-0.005
	(0.002)	(0.002)	(0.002)			
Primary 2004	0.025	0.062	0.037	0.017	0.045	0.008
	(0.005)	(0.003)	(0.004)			

Table 1: GOTV Example with Instrument as Potential Bias Amplifier. Results are displayed for estimates of the effect of the get-out-the-vote intervention on a number of pre-treatment outcomes thus creating placebo tests. Column 1 reveals that linear regression results suffer from bias due to selection on unobservables. Column 2 displays results from an extension of this analysis that could exacerbate the selection bias by including the indicator for the initial randomization, which in this case acts as an instrument. The third column presents the raw difference between columns 1 and 2. The final three columns decompose the bias into the constituent parts (see sections 2 and 3).

pared to 0.272 when regressing Y on Z and X. The net increase in bias is 97%. Here again, the major factor in the bias difference is bias amplification, α. Fixed effects are essentially pure bias amplifiers as evidenced by the fact there is a virtually no bias associated with omitting them (χ =0.001).

Results in Table 2, Panel A from other election years show substantively similar results for adding fixed effects to the model specification. Adding the fixed effects to the conditioning set of covariates leads to increased bias due to bias amplification.

Panel B of Table 2 presents the analysis where additional covariates are included in the specification. Again, refer back to model (5) to see the model specification. Y and Z are specified as in

Panel A. Here X_1 is a matrix of dummy variables for precinct and X_2 includes the prior election turnout indicators as in the Table 1, Panel B.

Results again show that fixed effects have amplified bias. While the amount of bias starts off lower for these models, the amplification factor is about the same, roughly doubling the bias of the estimate.

A. No covariates	OLS (SE)	FE (SE)	Diff (SE)	υ	α	χ
General 2004	0.138	0.272	0.134	0.137	0.135	0.001
	(0.004)	(0.004)	(0.004)			
General 2002	0.135	0.257	0.123	0.129	0.128	0.006
	(0.005)	(0.004)	(0.005)			
General 2000	0.131	0.257	0.126	0.129	0.128	0.002
	(0.005)	(0.004)	(0.004)			
Primary 2004	0.094	0.172	0.077	0.087	0.086	0.007
	(0.005)	(0.003)	(0.005)			
Primary 2002	0.093	0.174	0.081	0.087	0.086	0.006
	(0.004)	(0.003)	(0.004)			
Primary 2000	0.113	0.217	0.104	0.109	0.108	0.004
	(0.004)	(0.004)	(0.003)			
B. With covariates	**OLS (SE)**	**FE (SE)**	**Diff (SE)**	υ^*	α^*	χ^*
General 2004	0.017	0.037	0.02	0.018	0.019	-0.001
	(0.002)	(0.001)	(0.002)			
Primary 2004	0.025	0.037	0.011	0.018	0.019	0.007
	(0.005)	(0.002)	(0.005)			

Table 2: GOTV Example with Set of Fixed Effects as Potential Bias Amplifier. The columns are otherwise similar to those in Table 1.

6. CASE STUDY - THE EFFECT OF SELECTING A DISADVANTAGED VILLAGE COUNCIL PRESIDENT

In the previous case study we demonstrated a situation where bias amplification resulted from either adding an instrument or adding fixed effects to the conditioning set of covariates. This amplification occurred whether or not there were additional conditioning covariates specified in the model. In this section we consider a case study that repurposes data from another study (Dunning et al. 2013) to provide an example where bias amplification *per se* is not a major concern but where fixed effects nonetheless lead to a large increase in net bias as the bias due to υ is "unmasked."

The original paper examined the effect of having a village council president from a disadvantaged group (scheduled cast or scheduled tribe) on programmatic spending in India. In certain locations in India, council seats are reserved for disadvantaged groups on a rotating basis. Villages were assigned to have a reserved seat by first creating a list of councils within each district sorted by size of the population of the target disadvantaged group. Then, councils above a certain cutoff had their presidencies reserved for a disadvantaged group. In subsequent elections, the list was rotated so that a different set of villages had reserved seats. The original study capitalized on the list rotation scheme to conduct a quasi-experimental study. The analysis examined pairs of cities, one just above a cutoff for having a reserved presidency and one just below, and compared the difference in subsequent expenditures within pairs to examine whether councils with a reserved seat allocated more to public expenditure programs targeting the poor.

6.1. *Analysis and Results*

We reuse these data in a way not intended by the original study in order to induce confounding and study the resulting bias. We induce confounding by using the entire data set, not just the quasi-experimental pairs. Including data from all villages introduces confounding because villages higher on the list are not valid counterfactual cases for those further down given that they were sorted by the population of the disadvantaged groups.

Next, because outcome data exist for a time period *before* the assignment of the treatment, we were able once again to conduct a placebo test, whereby the effect of the treatment on the outcomes in a prior time period could be analyzed.[7] As above, since the treatment cannot affect outcomes

[7] We examined the "effect" of seats reserved in the 2007 election on outcomes from 2006. We also limited the data

in the past, the true value of the parameter is known to be zero. Estimates from the data can be compared to the true benchmark of zero and deviations from zero can be considered evidence of bias.

Our analysis compared models with and without fixed effects for district for each of a number of outcome measures. The estimates from the two models can be compared to see which is closer to the true parameter value of zero. If the fixed effects model is further from zero, then this is evidence that fixed effects cause a net increase in bias.

The outcome measures reflect seven government programs. Table 3 provides the names of the programs. Outcomes are measured in thousands of rupies for the first five outcomes and in number of latrines for the last two.

Reported in Panel A in the column labeled OLS in Table 3 is simply the naive non-adjusted OLS estimate of the effect of a reserved council presidency on expenditures. Again referring back to the model in (1), Y is an $(n \times 1)$ vector of expenditures (or number of latrines for the last two outcomes), Z is an $(n \times 1)$ vector of indicators for treatment assignment (reserved council presidency) and X is an $(n \times k)$ matrix of indicators of district (taluk).

In Panel A, Table 3 the column labeled OLS gives the estimated coefficient on Z when regressing Y on Z only. The column labeled FE gives the estimated coefficient on Z when regressing Y on Z and X.

Results in Panel A of Table 3 show that in three of seven cases (Ashraya, Latrines and Community Latrines), fixed effects appear to be moving estimates in the direction of zero. In two other cases (IAY Scheme and Ambedkar), the estimates are moving away from zero but only slightly so. In the case of MGNREGA, the result is a tossup with bias moving from 0.5 to -0.5 with the addition of the fixed effects.

In the case of the Water Infrastructure, however, the fixed effects estimate is much further from zero compared to the unadjusted estimate – 0.3 compared to -10.2. The bootstrapped standard

set to those villages that did not have a reserved presidency in 2005-2006 election years.

A. No Covariates	OLS (SE)	FE (SE)	Diff (SE)	υ	α	χ
Ashraya	3.4	0.4	-3.0	0.4	0.1	3.0
	(2.6)	(1.9)	(2.8)			
IAY	33.4	-34.0	-67.5	-30.0	-4.1	63.4
	(12.4)	(8.9)	(9.5)			
Ambedkar	-0.6	-0.8	-0.2	-0.7	-0.1	0.1
	(0.6)	(0.8)	(0.2)			
MGNREGA	0.5	-0.5	-1.0	-0.5	0.0	1.0
	(3.5)	(2.7)	(1.0)			
Water Infrastructure	0.3	-10.2	-10.5	-9.0	-1.2	9.3
	(3.9)	(4.0)	(2.6)			
Latrines	-12.3	-5.7	6.7	-5.0	-0.6	-7.3
	(5.8)	(5.3)	(4.0)			
Community Latrines	0.2	0.1	-0.2	0.0	0.0	0.2
	(0.2)	(0.1)	(0.1)			

B. With Covariates	OLS (SE)	FE (SE)	Diff (SE)	υ^*	α^*	χ^*
Ashraya	1.1	1.5	0.4	1.3	0.2	-0.3
	(1.7)	(2.2)	(1.8)			
IAY	-5.9	-10.0	-4.0	-8.9	-1.1	2.9
	(7.5)	(6.8)	(5.4)			
Ambedkar	-0.8	-1.6	-0.8	-1.5	-0.1	0.6
	(0.8)	(1.5)	(0.8)			
MGNREGA	-0.3	-0.6	-0.3	-0.6	0.0	0.3
	(3.6)	(3.2)	(0.7)			
Water Infrastructure	1.8	-4.1	-5.9	-3.6	-0.4	5.4
	(3.5)	(2.8)	(1.8)			
Latrines	-5.1	-1.8	3.4	-1.6	-0.2	-3.6
	(5.0)	(5.5)	(3.3)			
Community Latrines	0.1	0.2	0.1	0.2	0.0	-0.1
	(0.2)	(0.2)	(0.1)			

Table 3: Village Council Presidency Example with Fixed Effects

error suggests that this is a statistically significant difference.

In the last three columns of the table the observed biases are decomposed into constituent parts: the bias due to the unobserved confounder (υ), the bias due to amplification when controlling for X (α) and the bias due to omitting X from the conditioning set (χ). In decomposing this bias in the case of Water Infrastructure we can examine what is happening; the bias due to omitting X is roughly of the same magnitude as the bias due to omitting U but they have opposite signs. When neither X nor U are controlled for in the model, the two biases cancel. In this sense, χ can be said to be "good" bias which is masking υ in the unadjusted model. Bias amplification (α) plays a role, albeit a smaller one, accounting for about 9% (-0.6/-6.7) of the move away from zero when going from the unadjusted to the adjusted estimator.

Panel B of Table 3 shows the results for the model that includes several additional covariates in the conditioning set. Referring to model (5), Y and Z are defined as in Panel A. X_1 is a matrix of dummy variables for district. X_2 is an $(n \times 7)$ matrix of covariates including village expenditures for the year, village population, population of scheduled caste members, population of scheduled tribe members, size of the literate population, and size of the working population.

The values in the OLS column of Panel B, Table 3 are the estimated coefficient on Z when regressing Y on Z and X_2. The values in the FE column are the estimated coefficient on Z when regressing Y on Z, X_1 and X_2.

Results confirm the main finding for Water Infrastructure. Including fixed effects in the model unmasks the bias due to U, υ^*, making the total bias worse than when fixed effects are not included in the model.

Interestingly, results in Panel B also show that when controlling for these other covariates, X_2, fixed effects actually increase bias for 6 out of 7 outcomes compared to 2 out of 7 in Panel A. That additional covariates, X_2, can alter whether fixed effects help or hurt further complicates the question of what to include in the conditioning set for practitioners.

7. A SENSITIVITY ANALYSIS FRAMEWORK

For the case studies that we have examined, we have identified situations where adding an instrument or fixed effects to a set of conditioning variables increases bias. We can see this increase in bias because we have constructed these studies as placebo tests, whereby the true parameter value is known to be zero because the outcomes occurred before the treatment. However, our case studies provide little consolation to practitioners who do not know the true value of the parameter. The question we consider in this section is whether sensitivity analysis could be used to alert a practitioner to the potential for increases in bias that we have demonstrated.

Sensitivity analysis has been proposed as a way to visualize the potential for an unobserved confounder to bias results of an analysis (c.f. Imbens 2003; Rosenbaum and Rubin 1983; Clarke 2005, 2009). These approaches posit the attributes of an unobserved confounder, U, (for instance its association with treatment and outcome) that would be sufficient (in addition to observed confounders) to satisfy the selection on observable assumption. Then they calculate the amount of bias induced by failing to include U in the conditional set. Typically a full sensitivity analysis repeats this exercise across a range of possible attributes for U and the results can be visually displayed. If the estimated outcome changes very little except in the face of very extreme confounding by U, the results are said to be insensitive to omitted confounder bias. Similar attributes of observed covariates (for example their associations with treatment and outcome) can be used as benchmarks to help understand the range of plausible attribute values for "typical" covariates in that setting.

We modify for our purposes a new sensitivity analysis package available in R, treatSens (Carnegie et al. 2014b,a). The tool takes a dual-parameter approach similar to that of Imbens (2003). Similar to Imbens (2003) it is a likelihood-based method, but one that generates candidate values of the omitted confounder, U, drawing directly from the distribution of the confounder conditional on the observed data. Then the model to estimate the treatment effect is fit with the simulated confounder, $\dot{U}$.

For a given combination of values of the sensitivity parameters (the coefficients on U in the Y and Z models), ζ^z and ζ^y, an estimate of the treatment effect, τ, can be generated by first drawing candidate values of U, denoted $\dot{U}$, from the distribution implied by the sensitivity parameters and then estimating the parameters of the model regressing Y on Z, X_1, X_2 and $\dot{U}$ using OLS. Call the estimate of the coefficient on Z, $\widehat{\tau}(\zeta^z, \zeta^y)$. An average of this parameter estimate is taken across 20 draws of $\dot{U}$ to reduce the uncertainty associated with the random draws from the distribution of U. The algorithm proceeds by considering a range of possible values of ζ^z and ζ^y in a grid. Values of $\widehat{\tau}(\zeta^z, \zeta^y)$ can be computed for each cell in the grid. The values in the grid can be represented on a plot with axes ζ^z and ζ^y and contours drawn representing constant values of $\widehat{\tau}(\zeta^z, \zeta^y)$. The contours of the figure can be labeled with the associated value of $\widehat{\tau}(\zeta^z, \zeta^y)$.

Note that for a given contour all of the bias terms in our equations are identified. Therefore, we modify the sensitivity analysis currently available in the treatSens package to label each contour in the grid with the values of υ^*, α^*, and χ^*, which are themselves implied by the values of the sensitivity parameters, ζ^z and ζ^y, in addition to $\widehat{\tau}(\zeta^z, \zeta^y)$. Additionally, we place a contour demarcating the area in which the fixed effects would increase bias, rather than decreasing it. This modification allows the user to identify whether the areas of the parameter space where bias increases due to the addition of an additional variable (or set of variables) represent manifestations of the unobserved variable U that are plausible.

To calibrate the strength of the sensitivity parameters, we follow Imbens (2003) in plotting the coefficient estimates on the observed covariates, X_2, in the data. Given that the simulation proceeds assuming U has unit variance, covariate values, X_2, are standardized before estimating the coefficients. Confounders with negative partial associations with the outcome are reverse-coded so that they appear in the plot.

In the next section, we consider a sensitivity analysis plot for the voter turnout experiment. In Appendix C, we also present a figure for the data on village expenditures in India.

The sensitivity analysis for the GOTV outcomes, in Figure 7.1, examines the potential for fixed effects to bias the estimates for the effect of contact on General 2004 turnout, presented in Panel B of Table 2.

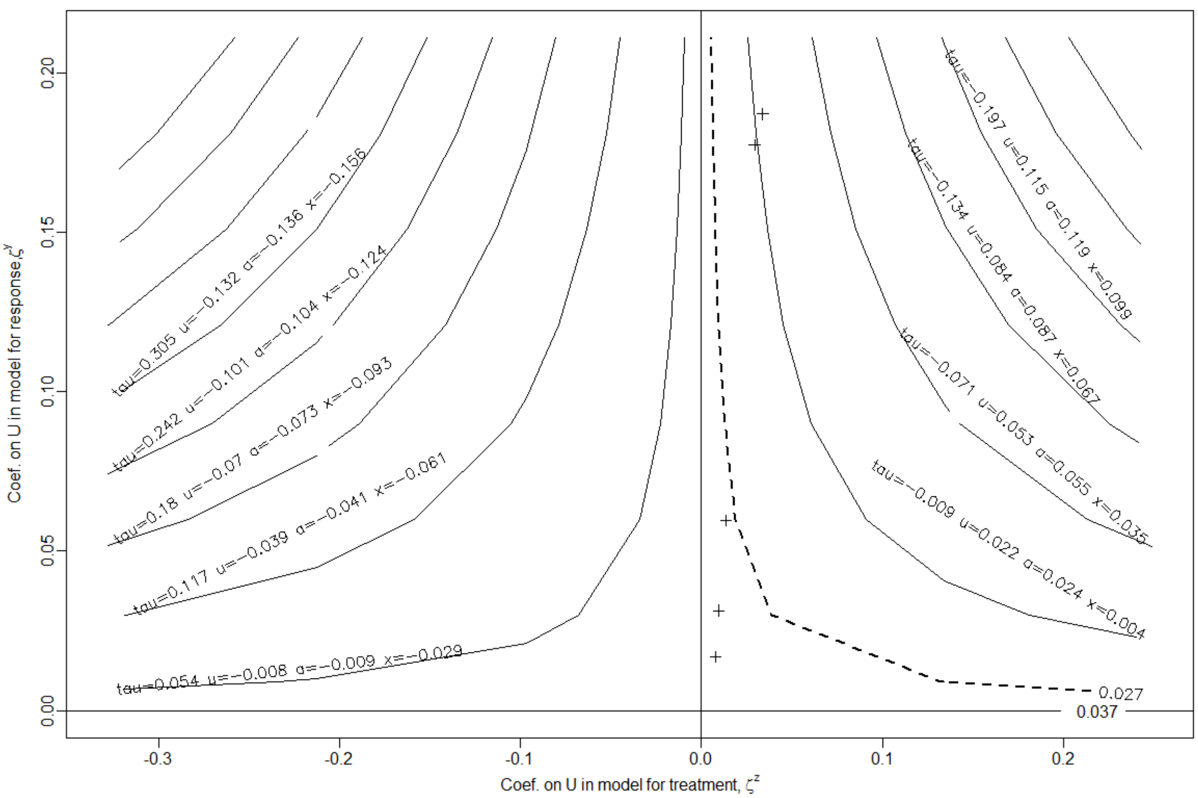

Figure 1: Sensitivity Plot For Gerber et. al. Data

In interpreting Figure 7.1, consider the point (0.1, 0.05). It falls approximately on the line labeled tau=-0.007. The figure implies that if $\zeta^z = 0.1$ and $\zeta^y = 0.05$, then the true effect would be about -0.007. The line also provides the decomposed bias $\upsilon = 0.021, \alpha = 0.023$ and $\chi = 0.003$. From this we can conclude that, if $\zeta^z = 0.1$ and $\zeta^y = 0.05$, then the bias of the estimator without fixed effects, $\upsilon + \chi = 0.024$, is smaller in magnitude than the bias when adjusting for fixed effects,

$\upsilon + \alpha = 0.044$. The figure also alerts us that the net amplification bias, α, is relatively large in this case being roughly 100% of the value of the omitted confounder bias, υ, throughout the figure. As a helpful summary, the dashed line represents the threshold separating the region where fixed effects are bias increasing from the region in which the fixed effects are bias reducing. Above and to the right of the line, fixed effects are bias increasing; for all other values of ζ^z and ζ^y the fixed effects are bias reducing.

The plus signs in the figure represent estimated coefficients on the (standardized) covariates, X_2, from the outcome and treatment models. As in Carnegie et al. (2014a) and Imbens (2003) we interpret the plus signs as providing benchmarks that help the researcher assess the plausibility of an omitted confounder with similar properties. For example the mark furthest from the origin, at about (0.03, 0.19), is plotting the coefficients on the (standardized) indicator of turnout in the 2000 general election. One interpretation is that it is not unlikely that the sensitivity parameters corresponding to the omitted confounder could have properties similar to that of the indicator for turnout in the 2000 general election. Reassuringly, Figure 7.1 indeed would have provided a researcher with a warning to be wary of fixed effects in this dataset.

In Appendix C, Figure 2, we examine a sensitivity plot for the Water Infrastructure outcome from the (Dunning et al. 2013) data.

8. DISCUSSION

We have discussed the ways in which additional control covariates can increase bias, including bias amplification and bias unmasking. In so doing, we identified a new special case of bias amplification, in particular, when fixed effects amplify bias. The canonical example of a (pure) bias amplifying covariate in the literature to date has been instruments. However, fixed effects can be pure bias amplifiers even though they do not act as instruments and even though they absorb heterogeneity in (and are causally related to) the outcome. We then presented a method of visualizing the conditions under which fixed effects are bias increasing (either via unmasking or amplification).

The bias formulas provided in this paper help us to better understand the circumstances under which covariates may act as bias amplifiers or bias unmaskers. Examining $\alpha \equiv \left(\frac{r^2_{Z|X}}{1-r^2_{Z|X}}\right)\upsilon$ provides some reassurance that amplification may not be a major concern in practice. It is only greater than the bias due to omitting U, υ, when $r^2_{Z|X} > \frac{1}{2}$. In words, X would have to account for more than half of the variability in the assignment mechanism for amplification to have the bias to be as as large as the bias due to the unobserved confounder.[8] Fortunately $r^2_{Z|X}$ is identified, a fact that should give us some idea of whether bias amplification should be a particular concern.

However, concern over the phenomenon of bias unmasking should perhaps rival concern over bias amplification. In the second case study, for example, the Water Infrastructure outcome reveals that the bias due to omitting fixed effects, χ, can be large, but of opposite sign and similar magnitude compared to the bias due to the unobserved confounder, υ. Omitting both the fixed effects and the unobserved confounder was preferable to adjusting for the fixed effects precisely because the two biases counterbalanced one another in the unadjusted estimate. In practice, a researcher is unlikely to know whether adjusting for covariates will unmask unobserved confounder bias. Similar observations have lead to somewhat pessimistic assessment of observational analysis, for example, in Clarke (2005) and Frisell et al. (2012) (but see also Clarke 2009).

Sensitivity analysis when considering unobserved confounders has been previously considered elsewhere (e.g. Clarke 2009; Carnegie et al. 2014a; Imbens 2003). We proposed a simple modification to sensitivity plots aimed at increasing the information available about potential for bias amplification. Plotting the bias decompositions (α, υ, and χ) on each of the contours may help practitioners to consider bias amplification.

While better study designs are always the best way to address concerns over the dangers caused

[8]In the case where X is a matrix of dummy variables for group, this condition is equivalent to saying that the intraclass correlation (ICC) is greater than 0.5.

by failing to control for all potential confounders in an observational studies, the reality is that many questions of interest are difficult or impossible to study using randomized experiments. In the absence of controlled or natural experiments we need more tools to help applied researchers make the best choices regarding how to perform their analyses. Thoughtful consideration about the potential for bias amplification *and unmasking* should be part of this process. We hope that the methodology presented in this paper can assist the researcher and makes these ideas more concrete.

REFERENCES

Angrist, J.D., Imbens, G. and Rubin, D. (1996). *Journal of the American Statistical Association* 91(434): 444-455.

Angrist, J.D. and Pischke, J. (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

Austin, P. Grootendorst, P. and Anderson, G. (2007) A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables Between Treated and Untreated Subjects: A Monte Carlo Study. *Statistics in Medicine* 26: 734-753.

Breen, R. K. Karlson, and A. Holm (2013). Total, direct, and indirect effects in logit and probit models. *Sociological Methods and Research*, 42(2), 164191.

Brookhart, M., Sturmer, T., Glynn, R., Rassen, J. and Schneeweiss, S. (2010). Confounding Control in Healthcare Database Research. *Medical Care* 48: S114-S120.

Bhattacharya, J. and Vogt, W. (2007). Do Instrumental Variables Belong in Propensity Scores? *NBER Working Paper* 343, National Bureau of Economic Research, MA.

Carnegie, NB, J Hill, and M Harada. (2014) Assessing sensitivity to unmeasured confounding using simulated potential confounders. (unpublished)

Carnegie, NB, J Hill, and M Harada. (2014) Package: TreatSens, http://www.R-project.org.

Clarke, K. A., (2005) The Phantom Menace. *Conflict Management and Peace Science* 22:341352.

Clarke, K. A., (2009) Return of the Phantom Menace. *Conflict Management and Peace Science* 26: 46-66.

Cole, SR, R W Platt, E. F. Schisterman, H. Chu, D. Westreich, D. Richardson, and C. Poole (2010). Illustrating bias due to conditioning on a collider. *Int J Epidemiol* 39(2): 417420.

D'Agostino, Jr. R. (1998). Propensity Score Methods for Bias Reduction in the Comparison of Treatment to Non-Randomized Control Group. *Statistics in Medicine* 17 314-316.

Ding, Peng and Miratrix, Luke. (2014). To Adjust or Not to Adjust? Sensitivity Analysis of M-bias and Butterfly-Bias. *Journal of Causal Inference* 2: 2-17.

Dunning, T. and Nilekani, J. (2013). Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils. *American Political Science Review* 107: 35-56.

Freedman, D. A. (2008). Randomization Does Not Justify Logistic Regression. *Statistical Science* 23(2): 237-249.

Frisell, T., Oberg, S., Kuja-Halkola, R., and Sjolander, A. (2012). Sibling Comparison Designs: Bias From Non-Shared Confounders and Measurement Error *Epidemiology* 23(5): 713-720.

Greene, WH. (2000). *Econometric Analysis* Prentice Hall (4th Edition).

Greenland, S. (2002). Quantifying biases in causal models: classical confounding vs. collider-stratification bias. *Epidemiology* 14: 300-306.

Hill, J. (2007a). Discussion of research using propensityscore matching: Comments on A critical appraisal of propensityscore matching in the medical literature between 1996 and 2003 by Peter Austin, *Statistics in Medicine* 27(12): 2055-2061

Hill, J. (2007b). ”Discussion of ’Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments’” *Journal of the American Statistical Association* 103(484): 1346–1350.

Heckman, J. and R. Robb. (1985). ”Alternative Methods for Estimating the Impact of Interventions” in J.J. Heckman and B. Singer (Eds.) *Longitudinal Analysis of Labor Market Data*: Cambridge University Press.

Heckman, J. and R. Robb. (1986). ”Alternative Methods for Solving the Problem of Selection bias in Evaluating the Impact of Treatments on Outcomes” in H. Wainer (Ed.) *Drawing Infereces from Self-selected Samples*: New Jersey: Lawrence Erlbaum Associates.

Imbens, Guido, W. (2003). Sensitivity to Exogeneity Assumptions in Program Evaluation. *Recent Advances in Econometric Methodology* 93(2): 126-132.

Lechner, M., (2001), Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in Lechner and Pfeiffer (Eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg, Physica.

Liu, W., M.A. Brookhart, S. Schneeweiss, X. Mi, and S. Setoguchi (2012). Implications of m-bias in epidemiologic studies: a simulation study. *American Journal of Epidemiology* 176: 938-948.

Myers J.A.,, J.A. Rassen, J.J. Gagne, K.F. Huybrechts, S. Schneeweiss, K.J. Rothman, M.M. Joffe, R.J. Glynn. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174(11):1213-22.

Pearl, Judea. (2000) *Causality* Cambridge.

Pearl J. (2009). Myth, confusion, and science in causal analysis. technical report.

Pearl, Judea. (2010). On a Class of Bias-Amplifying Variables that Endanger Effect Estimates. *Proceedings of UAI*, 417-424.

Pearl, J. (2011) Invited Commentary: Understanding Bias Amplification. *American Journal of Epidemiology* 174(11): 1223-1227.

Rosenbaum, P.R. (2002). *Observational Studies*, Springer.

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

Rubin, Donald (2002). Using Propensity Scores to Help Deign Observational Studies: Application to the Tobacco Litigation *Health Services and Outcomes Research Methodology* 2 169-188.

Rosenbaum, P. R. and Rubin, D. B. (1983), Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* 45: 212218

Shaw, D. R., D. P. Green, J. G. Gimpel, and A. S. Gerber. (2012) Do Robotic Calls from Credible Sources Influence Voter Turnout or Vote Choice? Evidence from a Randomized Field Experiment. *Journal of Political Marketing* 11 (4): 231-245.

Schisterman, E.F., Cole, S. R., and R. W. Platt. (2009). Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiology* 20:488-495.

Sjlander, A. (2009). Propensity scores and M-structures. Stat Med 28:141620.

Sobel, M.E., (2006). What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference. *Journal of the American Statistical Association* 101: 1398-1407.

Wooldridge, Jeffrey. (2009). Should Instrumental Variables Be Used as Matching Variables?. *Unpublished.*

Wyss, R., M. Lunt, M. A. Brookhart, R. J. Glynn, T. Strumer. (2014) *Journal of Causal Inference* 2(2): 131-146.

VanderWeele, T.J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction* Oxford University Press.

VanderWeele, T.J., and O.A. Arahc. (2011) Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Bias Formulas for Sensitivity Analysis. *Epidemiology*. 22(1): 4252.

A. OMITTED VARIABLE BIAS

To define the bias, start with a generic linear model,

$$Y = S\beta^s + O\beta^o + \epsilon^y, \tag{13}$$

where S and O are matrices of specified and omitted covariates, respectively. With respect to the error term, ϵ^y, assume $\mathrm{E}[\epsilon^y|S, O] = 0$.

Imagine the regression of Y on a set of covariates S only. This leads to the well known expression for omitted variable bias (for example see Greene 2000, p. 334)

$$\text{Bias}\left[\widehat{\beta}^s\right] = \mathrm{E}\left[S'S\right]^{-1} S'O\beta^o. \tag{14}$$

From this generic equation we can derive biases for particular sets of conditioning variables, S, under an assumed model.

To derive omitted variable bias, first collect variables into two groups: omitted variables, O, and included (specified) variables, S. Then we can write (in matrix notation) the general case

$Y = S\beta^s + O\beta^o + \epsilon^y$. Now substitute Y in $[S'S]^{-1} S'Y$ and take the expected value.

$$
\begin{aligned}
\mathrm{E}\left[\ \widehat{\beta}^s\ \right] &= \mathrm{E}\,[S'S]^{-1}\,S'Y \\
&= \mathrm{E}\,[S'S]^{-1}\,S'\left(S\beta^s + O\beta^o + \epsilon_y\right) \\
&= \beta^s + \mathrm{E}\,[S'S]^{-1}\,S'\left(O\beta^o + \epsilon_y\right) \\
&= \beta^s + \mathrm{E}\,[S'S]^{-1}\,S'O\beta^o
\end{aligned} \tag{15}
$$

The last line follows from the fact that ϵ_y is independent of S and O. Therefore, the bais is the last term,

$$
\text{Bias}\left[\widehat{\beta}^s\right] = \mathrm{E}\,[S'S]^{-1}\,S'O\beta^o. \tag{16}
$$

For the bias when conditioning on $S = [Z, X]$ and $O = [U]$, use the inverse of the partition matrix (cf. Greene 2000, section 2.6.3) to arrive at

$$
\begin{aligned}
&\text{Bias}\begin{bmatrix} \widehat{\tau} \\ \widehat{\beta}^y \end{bmatrix} \\
=&\mathrm{E}\begin{bmatrix} \left(Z'Z - Z'X\,[X'X]^{-1}\,X'Z\right)^{-1} & -\,[Z'Z]^{-1}\,Z'X\left(X'X - X'Z\,[Z'Z]\,Z'X\right)^{-1} \\ -\,[X'X]^{-1}\,X'Z\left(Z'Z - Z'X\,[X'X]^{-1}\,X'Z\right)^{-1} & \left(X'X - X'Z\,[Z'Z]^{-1}\,Z'X\right)^{-1} \end{bmatrix} \\
&\times\begin{bmatrix} Z'U \\ X'U \end{bmatrix}\zeta^y \\
=&\mathrm{E}\begin{bmatrix} \left(Z'Z - Z'X\,[X'X]^{-1}\,X'Z\right)^{-1} Z'U\zeta^y \\ -\,[X'X]^{-1}\,X'Z\left(Z'Z - Z'X\,[X'X]^{-1}\,X'Z\right)^{-1} Z'U\zeta^y \end{bmatrix}
\end{aligned} \tag{17}
$$

The last line follows from the fact that, by construction, $U \perp X$.

B. ILLUSTRATIVE EXAMPLE

In this section we provide a simple numerical example to illustrate these biases. Suppose a researcher has city level data and is interested in the effect of (standardized) per capita real income, Z, on (standardized) proportion voting for the legislative party in power, Y. Suppose the proportion voting for the party in power is also affected by whether or not the local mayor is a member of the the same party (a reverse coat-tails effect) such that

$$Y = \frac{1}{2}Z + \frac{1}{2}U + \epsilon_y$$
$$\epsilon_y \sim N\left(0, \frac{1}{2}\right).$$

where U is an indicator coded -1 if the mayor is not of the incumbent party and 1 if the mayor is of the incumbent party and ϵ_y represents idiosyncratic factors. For simplicity say half of mayors are members of the party in power.

Now in turn suppose the treatment, (standardized) per capita income, is affected by whether the mayor is the same party as the party in power in the legislature (because the legislature rewards districts with mayors of the same party with pork spending) and also by a development project aimed at increasing the incomes of the poor that was randomly assigned to half of the cities. The model for the treatment variable, Z (per capita income), is

$$Z = \frac{1}{2}X + \frac{1}{2}U + \epsilon_z$$
$$\epsilon_x \sim N\left(0, \frac{1}{\sqrt{2}}\right)$$

where X is an indicator of whether the development project took place in the district (coded -1 for not treated and 1 for treated).

Note that the example has been contrived such that $\mathrm{E}(Y) = \mathrm{E}(X) = \mathrm{E}(Z) = \mathrm{E}(U) = 0$ and

Variable	Measure	Scale
Y	Standardized proportion voting for legislative party in power	$\mu = 0, \sigma = 1$
Z	Standardized per capita income	$\mu = 0, \sigma = 1$
U	Mayor member of party in power	1 if yes, -1 if no
X	Development project in city	1 if yes, -1 if no

Table 4: Variables in illustrative example

$V(Y) = V(X) = V(Z) = V(U) = 1$. Note also that X is an instrument because it was randomly assigned to districts and only affects Y through its effect on Z.

Now suppose the researcher observes whether or not the development project occurs in each city but neglects to collect data on the party of the mayors. The researcher estimates the effect of (standardized) income per capita on (standardized) proportion voting for the political party in power in two ways. The first approach is to regress Y (proportion voting for party in power) on Z (per capita income). The second approach is to regress Y (proportion voting for party in power) on both Z (per capita income) and X (development project instrument).

We can compute the bias components for these specifications. The bias due to omitting X (development project instrument) is

$$
\begin{aligned}
\chi \equiv & \mathrm{E}\left[(Z'Z)^{-1} Z'X\beta^y \right] \\
= & \mathrm{E}\left[\mathrm{Cov}(Z, X) 0 \right] \\
= & 0
\end{aligned}
$$

as expected since X is an instrument. Meanwhile, the bias due to omitting U (mayor is of incumbent party) is

$$
\begin{aligned}
\upsilon \equiv & \mathrm{E}\left[\left(Z'Z\right)^{-1} Z'U\zeta^{y}\right] \\
= & \mathrm{E}\left[\mathrm{Cov}(Z,U)\frac{1}{2}\right] \\
= & \mathrm{E}\left[\mathrm{Cov}\left(\frac{1}{2}X+\frac{1}{2}U+\epsilon_z,U\right)\frac{1}{2}\right] \\
= & \frac{1}{4}V(U) \\
= & 0.25.
\end{aligned}
$$

The bias due to amplification is

$$
\begin{aligned}
\alpha \equiv & \left(\frac{r^2_{Z|X}}{1-r^2_{Z|X}}\right)\upsilon \\
= & \left(\frac{\mathrm{Cov}(X,Z)^2}{1-\mathrm{Cov}(X,Z)^2}\right)0.25 \\
= & \frac{1/4}{1-1/4}0.25 \\
= & 0.0833
\end{aligned}
$$

So the bias when omitting the instrumental variable X will be $(\chi+\upsilon)=(0+0.25)=0.25$ while the bias when including X in the conditioning set will be $(\alpha+\upsilon)=(0.0833+0.25)=0.333$. Thus in this case the unadjusted estimator is less biased than the estimator that includes the instrument in the conditioning set.

Why does this make sense intuitively? Omitting X is problematic because the development project affects per capita income in a way that in turn affects the proportion voting for the party in power; thus failing to control for X will create comparisons between units (cities) that only differ on income because of the development project (and therefore aren't truly comparable).

On the other hand, when we condition on X it sets up a comparisons within two groups:

those that received the development project and those that did not. Since X and U are negatively correlated conditional on Z (even though they are marginally uncorrelated), having a development project in the city makes is more likely that the mayor is a member of the party in power among precincts with higher per capita income and the reverse is true among precints with lower per capita income. Thus making comparisons within categories of X merely serves to accentuate the bias caused by omitting U.

The phenomenon of bias amplification is similar in the case of fixed effects in that conditioning on this additional variable sets up within-group comparisons that induce a negative relationship between U and X that exacerbates the bias.

C. SENSITIVITY ANALYSIS OF WATER OUTCOME

In the sensitivity plot, Figure 2, we examine the Water Infrastructure outcome. The interpretation of the plot is the same as in the case of the GOTV study. The preponderance of covariates whose benchmarking values fall in the bias-inducing region suggest that we should be concerned with including fixed effects in our analysis. Conducting sensitivity analysis on this data would have alerted the researcher for signs of potential trouble.

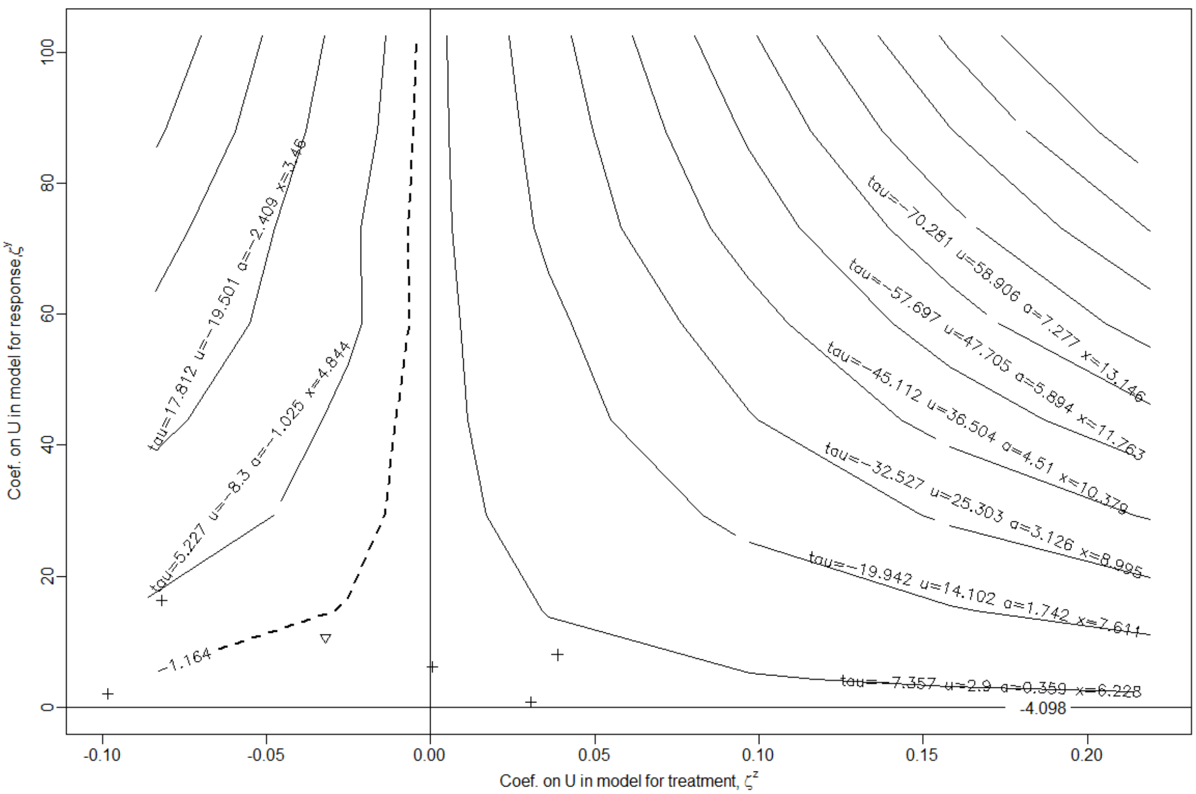

Figure 2: Sensitivity Plot of Water Outcome