# Learning from Biased Research Designs*

Andrew T. Little† Thomas B. Pepinsky‡

August 22, 2018

**Abstract**

Most contemporary empirical work in political science aims to learn about causal effects from research designs that may be subject to bias. We provide a Bayesian framework for understanding how researchers should approach the general problem of inferring causal effects from potentially biased research designs. The key to our approach is that both researchers and their audiences have prior beliefs about both causal effects and the degree and direction of bias. Once these priors are specified, what a rational researcher should learn from a potentially biased estimate can be derived from Bayes' rule. We apply this principle to explore when we should learn more or less from basic difference of means estimates, and then extend our analysis to speak to common modern designs intended to uncover causal effects.

*Many thanks to Neal Beck, Sean Gailmard, Andy Hall, and Cyrus Samii for comments and discussion.

†Assistant Professor of Political Science, UC Berkeley, andrew.little@berkeley.edu.

‡Associate Professor of Government, Cornell University, pepinsky@cornell.edu.

# 1 Introduction

Formal theories in political science frequently employ a learning process of the following form. An individual cares about some fact about the world, represented with a random variable $\delta$. She starts with a prior belief about $\delta$, and then observes a signal of $\delta$ which is contaminated by noise, represented with another random variable $\gamma$. A tractable way to formulate the problem is that the observed signal is equal to $\delta + \gamma$. Things become even more tractable when assuming $\delta$ and $\gamma$ are normally distributed (and independent). This problem is only marginally more complex if there is another (additive, and normally distributed) source of noise $\epsilon$, and so the signal observed is $\delta + \gamma + \epsilon$.

Most empirical work in contemporary political science involves trying to learn about a treatment effect or causal effect $\delta$. This is frequently done by calculating a difference of means or estimating a regression which returns an estimate of the form $\delta + \gamma + \epsilon$, where $\gamma$ is the bias in the estimate, and, when appropriate, $\epsilon$ is the sampling error. Various assumptions of normality (typically about $\epsilon$) are invoked in this setting as well. In the wake of the "credibility revolution" in the social sciences (Angrist and Pischke, 2010), empirical research has come to depend on developing research designs where the bias term is small and the sampling error has mean zero and a variance which can be estimated from the data. Arguments about research design and the "credibility" of estimates of causal effects frequently center around whether assumptions about the bias term are valid. These arguments can be substantive ("here is a plausible confounding variable in your regression") or statistical ("here is why the research design does or does not effectively deflect concerns about bias").

Our central contention in this paper is we can reconcile the practice of learning about causal effects in applied empirical work with formal theories of learning in political science. We propose that, from the perspective of the model in the first paragraph, we may represent argument and disagreement about research design as scholars (either truly or instrumentally) holding different

prior beliefs about $\gamma$. The way we think about learning about causal effects in empirical work can therefore be grounded as forming joint posterior beliefs about parameters of interest and bias, just as actors in our formal theories of learning would do. Some modes of argument about empirical work make sense and can be clarified with this lens. Others become less defensible.

## 2 A Motivating Example

To fix ideas, consider researcher who seeks to investigate the effect of access to public transportation on support for government spending. This is interesting because the researcher suspects that if it were possible to increase the availability of public transportation, it would shift mass attitudes about government spending, thereby creating a constituency that supports the spending required to maintain public transportation. It is hard to assign people randomly to use public transportation, and the researcher suspects that a priming experiment or relatedly artificial manipulation is too unrealistic to be useful. So, she proceeds with observational data, collecting data on individuals' support for government spending as well as a binary variable $D$ that captures whether each individual uses public transportation. She then runs a standard regression of $Y = \delta D + \epsilon$.

Under the assumption that $Cov(D, \epsilon) = 0$, this regression will (in large samples) identify the true causal effect of the use of public transportation on support for government spending: $\hat{\delta} = \delta$. However, that is an implausible assumption. There are good reasons to suspect that the type of people who support government spending are also likely to be the types of people who use public transportation the first place. If so, the estimate $\hat{\delta}$ does not represent the causal effect of public transportation on individual attitudes. It represents the sum of that effect plus the effect which can be attributed to non-random selection. This matters for the researcher because she would not conclude that increasing public transportation would change mass attitudes towards government spending if only those people who favor government spending will use it.

Call a latent variable that represents that selection process $U$. If the researcher could observe

$U$ she would condition on it in a regression of $Y = \delta D + \beta U + \epsilon$. She knows, however, that by a standard calculation calculation of omitted variable bias she can represent the naive estimate as $\hat{\delta} = \delta + \beta \frac{Cov(D,U)}{Var(D)}$; that is, as the sum of the true effect of using public transportation plus the effect attributable to selection weighted by the relationship between selection and treatment. Relabeling $\beta \frac{Cov(D,U)}{Var(D)} = \gamma$, if either selection is orthogonal to treatment ($Cov(D, U) = 0$) or there are no selection effects ($\beta = 0$) then $\gamma = 0$ and $\hat{\delta} = \delta$.

Another way to describe our researcher's concern about identifying the causal effect of taking public transportation is that her prior belief about $\gamma$ places high probability on values meaningfully far from zero. If her main concern is that people who take public transportation tend to like higher government spending for non-causal reasons, then her prior should place more probability on positive values of $\gamma$.

Once we specify a prior belief about the joint distribution of $\delta$ and $\gamma$, it is straightforward in principle (if not always in computation) to use Bayes' rule to determine what our posterior beliefs about the true causal effect $\delta$ change, given those prior beliefs, if we observe a particular estimate $\hat{\delta}$. But to make sense of what we can learn, we need a model of learning itself that begins with prior beliefs about $\delta$ and $\gamma$.

## 3 Literature

Before proceeding, we review existing approaches to confronting the problem of learning about $\delta$ from observational data. This will allow us to distinguish our approach from current practice.

One way to learn about causal effects in observational research, which might be termed the "head-in-sand" approach, is to proceed as if there is no bias at all. In contemporary practice, this may take the form of adding a series of control variables to the regression and arguing that these are likely to take care of any causal identification concerns. Such practice is already subject to voluminous criticism in the empirical social sciences, so we do not address it here. Another approach,

which might be termed the "staring-at-the-sun" approach, is to impose complete agnosticism about the value of $\gamma$, in which case we can learn nothing about $\delta$ from observational data.

The modal contemporary approach is in contemporary political science and applied microeconomics focuses on design. The general approach is to identify conditions under which the assumption that $\gamma = 0$ is "credible" (see Angrist and Pischke 2010). Experiments ensure that $\gamma = 0$ because the assignment mechanism governing $D$—randomization—is known to be unrelated (in expectation) to any plausible confounders. In the context of our motivating example, where an experiment is infeasible, a credibility-based empirical approach may involve identifying subpopulations where it is known that individuals do not use public transportation because of their prior opinions about it, or where the assignment of individuals to use public transportation is governed by a known process, or a related strategy. The result is an estimate of a local average treatment effect (LATE), which is an unbiased estimate of $\delta$ for a particular subpopulation on the assumption that the research design is credible. Generalizing from the LATE to the population ATE can be difficult (Aronow and Carnegie, 2013), a point we address in section 6.2.

A different approach is inspired by Manski (1995). Rather than attempt to point-identify the parameter $\delta$, one may uncover an interval or range of values in which $\delta$ may lie based on minimal assumptions about unobserved features of the data (in this case, the value of $\gamma$). This "partial identification" approach can be valuable when it yields results that are substantively useful even if the precise value $\delta$ remains unknown; for example, the range of $\delta : [\underline{\delta}, \overline{\delta}]$ may be strictly positive. Common sources of leverage that can place bounds on the identification region are the distributional characteristics of the variables themselves, or prior information about the logically possible range of $\delta$. Alone or together, these may be used to derive bounds on the range of values that $\delta$ may take without making any assumptions about $\gamma$.

A third approach examines the sensitivity of $\hat{\delta}$ to various assumptions about $\gamma$. Rosenbaum (2002) illustrates how to explore how $\hat{\delta}$ changes when allowing the probability of treatment assignment to differ across matched cases of $D$. If large hypothetical differences in treatment assignment

probabilities yield small changes in $\hat{\delta}$ or fail to render it statistically indistinguishable from zero, then one may use $\hat{\delta}$ to learn about $\delta$ even without precise information about $\gamma$. Related approaches from Altonji, Elder and Taber (2005) and Oster (2017) proceed with a different assumption: that one may use the relationship between the treatment and observed confounders to approximate the relationship between the treatment and unobserved confounders. Under that assumption, one may also construct sensitivity tests that mimic the possible effects of $\gamma$ on $\hat{\delta}$ without any further information about $\gamma$.

These three approaches each confront the problem of unknown $\gamma$ in different ways: by finding a way to ensure that $\gamma$ is zero (the credibility approach), to see what can be learned without making reference to $\gamma$ (the partial identification approach), and to explore the sensitivity of results to various values of $\gamma$ (the bounds approach).

Broadly speaking, our model provide a synthetic way to think about these approaches. We show how any prior distribution on the relationship between the treatment effect, bias, and observed (local) treatment effects should map on to a posterior belief about the treatment effect. While we propose no new methods to design studies to provide unbiased or precise estimates of treatment effects, our approach highlights how we should be able to learn (and how much we should learn) from imperfect designs.

The closest precursor to this paper is Gerber, Green and Kaplan (2014), who use a similar approach to argue that if the prior beliefs about bias are completely uninformative, we should learn nothing from observational studies. That conclusion justifies the "staring-at-the-sun" approach to observational data described previously. We discuss the areas of overlap in between our results and theirs in Section 5, but here we highlight several important differences. First, we allow for more general prior beliefs, both by allowing for correlation between the bias and treatment priors when using normal distributions, and presenting results without the assumption of normality. Second, we focus more on the case where the prior beliefs do not have infinite variance, and argue extensively why this is a more appropriate assumption. Third, we show how the approach can be applied

to better understand some of the main contemporary classes of observational research design not addressed in Gerber, Green and Kaplan (2014).

## 4 General Setup: Binary Treatment, No Sampling Error

Consider a standard binary treatment potential outcomes setup. Let $\delta = \mathbb{E}[Y^1 - Y^0|D = 1]$ be the true average treatment effect on the treated (ATET) in a population (superscripts referring to potential outcomes, $D$ indicates treatment status). Let $\hat{\delta} = \mathbb{E}[Y_1^1|D = 1] - E[Y_0^0|D = 0]$ be the standard difference of means estimator (subscripts referring to actual treatment status). By a standard calculation, the difference of means estimate can be written:

$$\hat{\delta} = \delta + \gamma$$

where $\gamma = E[Y^0|D = 1] - E[Y^0|D = 0]$ is the selection bias.

Suppose we start with a join prior $f(\delta, \gamma)$ on the ATE and the bias. Upon observing $\hat{\delta}$, the researcher learns that the true $(\delta, \gamma)$ lies on the ridge given by $\hat{\delta} = \delta + \gamma$. So, the posterior marginal belief about $\delta$ is given by integrating out $\gamma$:

$$f_{\delta|\hat{\delta}}(\delta|\hat{\delta}) = \int_{\gamma} \frac{f(\hat{\delta} - \gamma, \gamma)}{\int_{(\delta,\gamma):\delta+\gamma=\hat{\delta}} f(\delta, \gamma) d(\delta, \gamma)} d\gamma$$

If we add sampling error to the mix with prior joint density $f(\delta, \gamma, \epsilon)$, the only change is to integrate over both sources of noise:

$$f_{\delta|\hat{\delta}}(\delta|\hat{\delta}) = \int_{\gamma,\epsilon} \frac{f(\hat{\delta} - \gamma - \epsilon, \gamma, \epsilon)}{\int_{(\delta,\gamma,\epsilon):\delta+\gamma+\epsilon=\hat{\delta}} f(\delta, \gamma, \epsilon) d(\delta, \gamma, \epsilon)} d(\gamma, \epsilon)$$

**Yes, You Have Priors**

Scholars may be uncomfortable specifying prior beliefs about treatment effects, let alone about their bias. The notion that we have prior beliefs about the main parameters of interest in our models – here, the treatment effect – *should* be relatively uncontroversial, and is a central part of any Bayesian inference.[1] However, because our approach hinges on researchers having priors even if they do not acknowledge them, it makes sense to address this point in more detail.

In practice – in the sense of "what people tend to do when estimating statistical models" – empirical researchers using Bayesian techniques usually employ "non-informative" (typically, high variance) and "neutral" (typically, mean $0$) priors. However, the way scholars actually discuss research reveals that they do have informative (and often non-neutral) priors. Observe that scholars frequently make statements like "that estimate seems reasonable" or "I don't believe the treatment effect could really be that large." Such statements are impossible without prior beliefs about the treatment effect. And when researchers are asked to make predictions about experimental effects, they can do so and express a degree of confidence in their predictions (DellaVigna and Pope, 2016). Further, these predictions are correlated with the observed results, and are more accurate for those who express higher confidence (i.e., a prior with lower variance).

If this is not persuasive that scholars almost always have meaningful priors about causal estimates, consider the converse. Suppose a scholar has conducted an airtight study to identify a treatment effect on a question you care about. In this context, to have no prior belief about the treatment effect is to say that *no result – hugely positive, hugely negative, or zero – would be more or less surprising to you*. We claim that there is virtually no realm of empirical social science research in which scholars are so agnostic about such a causal effect. Researchers have prior beliefs about treatment effects.

But what about prior beliefs about bias? One can represent complete ignorance about bias

[1] See also Gill and Walker (2005) for a discussion of how to construct statements about parts of a distribution to a full prior.

by examining the limiting case as the prior approaches an improper uniform prior (which is what many results in Gerber, Green and Kaplan 2014 do). However, the way that scholars debate the interpretation of empirical studies indicates they *do* have beliefs about the likely direction and magnitude of bias. To say in a seminar "I'm worried that your estimate is biased because of (insert confound/sampling/design issue)" is to say "I have a prior belief that $\gamma$ is more likely to be positive (or negative)." Claiming to be fully agnostic about the bias in estimates may be a useful conservative benchmark, but is clearly inconsistent with how scholars actually debate the merits of different research designs.

Still, one may be uncomfortable translating these prior beliefs into a specific prior distribution Gill and Walker (2005). This objection is analogous to the critique of theoretical models of learning more generally, where one could say "Even I, political scientist, cannot conjure up the mean and variance of my prior belief about (say) what economic growth will be under a proposed policy, so how could an even less informed voter do this?" The common response to this objection is that these models do not require voters to think literally about such prior distributions, just to learn more or less as if they do have priors approximated by those we assume for the purpose of analytical tractability. Similarly, scholars who adhere to the principle that Bayesian learning is optimal *should* learn about treatment effects and bias *as if* they have specific prior beliefs, even if they are uncomfortable directly articulating them.

Further, one can view the procedures that we outline below as equivalent to saying "if one held prior belief $f$, this is the resulting belief after observing the data." If scholars are "uncertain about their prior", they can examine how different priors would map to different conclusions. Our procedures below provide structure for such how to do this, and also yield insights that run contrary to what many researchers appear to believe we can learn from imperfect research designs.

**What Does it Mean to Learn?**

Once we agree that researchers have prior beliefs, we can be more precise about what it means to "learn" from a potentially biased study. At a high level of abstraction, we can say that learning happens whenever the posterior belief about the treatment effect $f_{\delta|\hat{\delta}}(\delta|\hat{\delta})$ is different from the prior marginal density $f(\delta)$. Unless the difference of means is independent of the true treatment effect, there will always be difference between prior and posterior beliefs—and thus, there will be learning.

To make more concrete statements about the degree to which we learn about a causal effect from a potentially biased study, we focus on learning in two numbered senses.

1. To measure *first-moment* learning, we study how much the mean of the posterior belief about $\delta$ changes as a function of $\hat{\delta}$. When our belief about $\delta$ changes rapidly as the difference of means decreases or increases, then we can say that we learn a great deal from a research design.

2. To measure *second-moment* learning, we study how much the variance of the posterior belief about $\delta$ decreases after observing $\hat{\delta}$.

Each of these notions of learning will prove convenient for certain results or calculations, but there is a strong and general relationship between the two. This follows immediately from the law of total variance, which can be stated in our context as:

$$\textit{Var}[\delta] - \mathbb{E}[Var[\delta|\hat{\delta}]] = \textit{Var}[\mathbb{E}[\delta|\hat{\delta}]] \tag{1}$$

That is, if we formalize first-moment learning as the variance in the posterior mean about $\delta$ (the right-hand side of Equation 1), this is always equal to the average decrease in the variance of the belief about $\delta$ (the left-hand side of Equation 1), which is a formalization of second-moment learning.

**The Unit Principle**

Finally, before placing any distributional assumptions on the prior, we clarify one aspect of how one should learn from empirical studies that is so simple as to be almost embarrassing to write. One way to think about first-moment learning is to ask "how different would my resulting average beliefs be if the estimator returned a different answer?" A formal statement of the claim that we (first-moment) learn nothing about a treatment effect from a particular research design is that mean of our posterior belief is invariant in the result. Formally, $\frac{\partial \mathbb{E}[\delta|\hat{\delta}]}{\partial \hat{\delta}} = 0$.

More generally, a question we might ask is "how do our beliefs about the treatment effect and bias changes when the difference of means increases?" In terms of how the mean of our beliefs about these variables, a property of this answer is immediate from the linearity of expectations and derivatives:

**Theorem 1.** ***(The Unit Principle)****: Suppose* $\hat{\delta} = \delta + \gamma$*. Then* $\frac{\partial \mathbb{E}[\delta|\hat{\delta}]}{\partial \hat{\delta}} + \frac{\partial \mathbb{E}[\gamma|\hat{\delta}]}{\partial \hat{\delta}} = 1$

**Proof**

$$\frac{\partial \mathbb{E}[\delta|\hat{\delta}]}{\partial \hat{\delta}} + \frac{\partial \mathbb{E}[\gamma|\hat{\delta}]}{\partial \hat{\delta}} = \frac{\partial \mathbb{E}[\delta + \gamma|\hat{\delta}]}{\partial \hat{\delta}} = \frac{\partial \mathbb{E}[\hat{\delta}|\hat{\delta}]}{\partial \hat{\delta}} = 1 \quad \blacksquare \tag{2}$$

So, a unit increase in the difference of means must be "divided" between an increase in the mean belief about the treatment effect and a mean belief about bias. The details of how the updating is divided into changing beliefs about treatment and bias will depend on the prior beliefs; our discussion below will show precisely how this happens. In special cases, one may update purely on the treatment effect ($\frac{\partial \mathbb{E}[\delta|\hat{\delta}]}{\partial \hat{\delta}} = 1$ and $\frac{\partial \mathbb{E}[\gamma|\hat{\delta}]}{\partial \hat{\delta}} = 0$, or a "credible estimate") or only on the bias ($\frac{\partial \mathbb{E}[\delta|\hat{\delta}]}{\partial \hat{\delta}} = 0$ and $\frac{\partial \mathbb{E}[\gamma|\hat{\delta}]}{\partial \hat{\delta}} = 1$, or "we learn nothing"). We will also see cases where the slope on one update is *greater* than one and the other is negative. But the unit principle is inviolable.

The unit principle also extends to the case with sampling error. If $\hat{\delta} = \delta + \gamma + \epsilon$, then $\frac{\partial \mathbb{E}[\delta|\hat{\delta}]}{\partial \hat{\delta}} + \frac{\partial \mathbb{E}[\gamma|\hat{\delta}]}{\partial \hat{\delta}} + \frac{\partial \mathbb{E}[\epsilon|\hat{\delta}]}{\partial \hat{\delta}} = 1$. We will entertain this case as an extension of our main results below.

## 5 A Model of Learning with Normal Distributions

To make more concrete statements about how one should learn from imperfect research designs, we need to place more structure on prior beliefs about $\delta$ (the treatment effect) and $\gamma$ (bias). A natural place to start is to assume that these two random variables are drawn from a multivariate normal distribution. Consider the case with no sampling error, and let the prior distribution on $(\delta, \gamma)$ be a multivariate normal with mean vector $(\mu_\delta, \mu_\gamma)$. The prior variances of the individual variables are $\sigma_\delta^2$ and $\sigma_\gamma^2$, and the covariance is $\rho\sigma_\delta\sigma_\gamma$, where $\rho$ is the correlation between the prior belief about bias and treatment. We provide some theoretical intuitions for when this correlation might be non-zero in Section 6.5.

The researcher's prior is that the difference in means will be $\mu_\delta+\mu_\gamma$. In general, upon observing a larger difference of means than $\mu_\delta + \mu_\gamma$ she will update positively on *both* $\delta$ *and* $\gamma$. Upon observing a smaller difference of means, she will generally update negatively.

By a standard property of multivariate normal distributions, the posterior belief about $(\delta, \gamma)$ upon observing $\hat{\delta}$ is normally distributed with means

$$\overline{\mu}_\delta = \mu_\delta + m_\delta(\hat{\delta} - \mu_\delta - \mu_\gamma) \tag{3}$$

$$\overline{\mu}_\gamma = \mu_\gamma + m_\gamma(\hat{\delta} - \mu_\delta - \mu_\gamma) \tag{4}$$

where

$$m_\delta = \frac{Cov(\delta, \hat{\delta})}{Var(\hat{\delta})} = \frac{\mathbb{E}[(\delta - \mu_\delta)(\delta - \mu_\delta + \gamma - \mu_\gamma)]}{\mathbb{E}[(\delta - \mu_\delta + \gamma - \mu_\gamma)(\delta - \mu_\delta + \gamma - \mu_\gamma)]} = \frac{\sigma_\delta^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta s_\gamma + \sigma_\gamma^2}$$
$$m_\gamma = \frac{Cov(\gamma, \hat{\delta})}{Var(\hat{\delta})} = \frac{\sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta s_\gamma + \sigma_\gamma^2}.$$

(See Appendix A for the full derivation.) For both variables, the updated belief is equal to the prior plus a fraction times the difference between the observed treatment effect and what the researcher

expects in the prior ($\mu_\delta + \mu_\gamma$). So, we can interpret this fraction as the magnitude of updating the point estimate on that variable. When $m_\delta > 0$ – i.e., when the covariance between $\delta$ and $\hat{\delta}$ is positive – the researcher has a more positive belief about the true treatment effect upon observing a higher difference of means.

The variances in the posterior distribution are:

$$\overline{\sigma}^2_\delta = \overline{\sigma}^2_\gamma = \sigma^2_\delta - \frac{Cov(\delta, \hat{\delta})^2}{Var(\hat{\delta})} = \frac{(1-\rho^2)\sigma^2_\delta \sigma^2_\gamma}{\sigma^2_\delta + 2\rho\sigma_\delta\sigma_\gamma + \sigma^2_\gamma} \tag{5}$$

### 5.1 Discussion

The expressions above are just formal representations—given particular assumptions about how we can represent prior beliefs—of how we ought to learn about unknown parameters given the data that we observe and the beliefs that we hold. In general, the results fit standard intuitions about how we ought to learn about those parameters. Closer inspection, moreover, reveals some usefully non-obvious results about the conditions under which "learning" happens even given uncertainty about these unknown parameters.

We begin by relating those theoretical results to some basic intuitions about how we learn from new data. Suppose the extreme case that we already knew the treatment effect $\delta$ with certainty, and then were provided with results from a research design in which there may be some form of bias. In that case, because we are certain about the treatment effect, we ought to put no weight on the results from a study with unknown bias. In our framework, complete certainty means that there is no variance in the prior belief about $\delta$. Formally, as $\sigma_\delta \to 0$, $\overline{\sigma}^2_\delta \to 0$ and $m_\delta \to 0$. Accordingly, the second term in Equation 3 reduces to zero, and $\overline{\mu}_\delta = \mu_\delta$. When we believe that we know the treatment effect with certainty prior to seeing the data, then regardless of the possible magnitude of the bias term, our posterior belief is identical to our prior belief; we put no weight on—and we learn nothing from—the results of the study with bias.[2]

[2]That $\overline{\mu}_\delta = \mu_\delta$ means that we "first-moment learn" nothing from this research design. It is also true that we

Now consider another case extreme case, in which we are certain about the magnitude of the bias parameter $\gamma$ for a particular research design. Intuitively, then, we ought to be able to learn about the true treatment effect from the biased estimate because we can simply back out the true effect by subtracting the known bias from the biased estimate. In our framework, we represent certainty as $\sigma_\gamma \to 0$, in which case $m_\delta \to 1$ and $\overline{\sigma}^2_\delta \to 0$. Thus, Equation 3 reduces to $\mu_\delta + \hat{\delta} - \mu_\delta - \mu_\gamma = \hat{\delta} - \mu_\gamma$, exactly the condition under which we can uncover the true treatment effect by subtracting the bias term from the biased estimate. Observe as well that an unbiased research design is a special case of this condition, where $\sigma_\gamma = 0$ and $\mu_\gamma = 0$ (meaning that our prior belief is certain that the bias term is exactly zero). In that case, our posterior estimate of the true treatment effect is simply the estimated treatment effect $\hat{\delta}$.

These results comport intuitively with what we expect about how we should learn from biased research designs. If we already know the parameter of interest we learn nothing from a biased research design. If we already know the bias exactly we can adjust our inferences to account for that bias. But what if neither of those conditions hold? It turns out that under very general conditions we can still learn something about our treatment effect of interest. Framed in terms of second-moment learning, an important consequence of (5) is:

**Proposition 2.** ***If there are priors, then there is learning.*** *If $\sigma_\delta$ and $\sigma_\gamma$ are both finite, then $\overline{\sigma}^2_\delta < \sigma^2_\delta$.*

**Proof** Rewriting the posterior variance

$$\overline{\sigma}^2_\delta = \sigma^2_\delta \frac{(1-\rho^2)\sigma^2_\gamma}{\sigma^2_\delta + 2\rho\sigma_\delta\sigma_\gamma + \sigma^2_\gamma} < \sigma^2_\delta \frac{\sigma^2_\gamma}{\sigma^2_\delta + 2\rho\sigma_\delta\sigma_\gamma + \sigma^2_\gamma} < \sigma^2_\delta \tag{6}$$

Proposition 2 establishes a general result that learning should happen, even from a research design subject to potentially massive bias, just so long as the prior beliefs of both the treatment effect and the bias term have finite variances. Learning here refers to changes in the posterior beliefs

"second-moment learn" nothing from this design, because $\overline{\sigma}^2_\delta \to 0$ in Equation 5.

about the variance of $\delta$ given $\hat{\delta}$, or second-moment learning—posterior estimates are always more precise than prior beliefs. What this means in practice is that, perhaps surprisingly, so long as the researcher holds a proper prior belief on $\delta$ and $\gamma$ she will have more precise beliefs about $\delta$ from seeing $\hat{\delta}$ from any research design than she would from not seeing $\hat{\delta}$. If, on the other hand, the researcher is unwilling to place any information in the prior belief about $\sigma_\gamma$ or $\sigma_\delta$, it follows that $\sigma_\gamma, \sigma_\delta \to \infty$ and as a result that the posterior variances $\overline{\sigma}^2_\delta, \overline{\sigma}^2_\gamma \to \infty$. In this case, the researcher's beliefs about $\delta$ is no more precise upon observing any $\hat{\delta}$.[3]

From that baseline result that second-moment learning always happens so long as we have proper priors, we can explore related cases where the researcher has proper priors over either the treatment effect or the bias term but not both. For any given level of uncertainty in the prior belief about the bias term $\gamma$, with complete uncertainty about $\delta$ ($\sigma_\delta \to \infty$, i.e. there is no prior information about $\delta$) then $m_\delta \to 1$. The posterior distribution about $\delta$ in this case is normally distributed with mean $\hat{\delta}$ and variance $(1-\rho^2)\sigma^2_\gamma$. In this case, the researcher updates strongly her *mean* estimate about the true treatment effect $\delta$ given the potentially biased estimate $\hat{\delta}$, but if $\sigma^2_\gamma$ is high (and the prior beliefs are not too strongly correlated), she can remain quite uncertain about the treatment effect. Conversely, for any given level of uncertainty in the prior belief about the treatment effect $\delta$, with complete uncertainty about $\gamma$ ($\sigma_\gamma \to \infty$, i.e. there is no prior information about the bias term), then $m_\delta \to 0$. The posterior distribution about $\delta$ in this case is normally distributed with mean $0$ and variance $(1-\rho^2)\sigma^2_\delta$. Comparing these two cases, extreme uncertainty about bias leads to an inability to update on the treatment effect; this is the main point of Gerber, Green and Kaplan (2014), who argue that learning from observational research—where the bias term is unknown, and represented as having infinite variance—is an "illusion."

With finite variances on both $\delta$ and $\gamma$, we can push our conclusions further. Recall that the covariance in the prior distributions of $\delta$ and $\gamma$ is given by $\rho\sigma_\delta\sigma_\gamma$. Where $\rho = 0$, the prior dis-

[3]There is still an update on the point estimate of $\delta$, corresponding to what we term first-moment learning, but this depends on the relative rates of convergence.

tributions of $\delta$ and $\gamma$ are independent of one another. In particular case, our framework reduces to a standard Bayesian learning problem where the observed treatment effect is equal to the real treatment effect plus noise (in our case, bias). Formally, $m_\delta = \frac{\sigma_\gamma^2}{\sigma_\gamma^2+\sigma_\delta^2} = \frac{\sigma_\delta^{-2}}{\sigma_\gamma^{-2}+\sigma_\delta^{-2}}$. Accordingly, the posterior mean of $\delta$ is a weighted average of the prior and the (de-meaned) signal, with weights are equal to the precisions of the prior and the signal. The more precise our prior belief about $\delta$ is, the less it changes in response to $\hat{\delta}$.

Are there conditions under which observing a higher value of $\hat{\delta}$ must lead us to increase our posterior estimate of $\delta$? Assuming a proper prior on $\gamma$, so long as it is the case that $\rho > -\frac{\sigma_\delta}{\sigma_\gamma}$ then it will also be true that $m_\delta > 0$, and as a consequence our posterior mean estimate of $\delta$ is increases even with the biased $\hat{\delta}$. An intuitive reason why this would not hold if the researcher is highly uncertain about $\gamma$ and also believes that $\gamma$ and $\delta$ are strongly negatively correlated. Since the variance in $\hat{\delta}$ mainly driven by variance in $\gamma$, observing a higher $\hat{\delta}$ primarily makes the researcher think $\gamma$ is higher. And if she has a strong prior belief that $\delta$ and $\gamma$ are negatively correlated, this can push the belief about $\delta$ downwards.

Finally, it bears emphasis that the coefficients which determine how increases in observed treatment effects – $m_\delta$ and $m_\gamma$ – are not a function of $\hat{\delta}$. This implies that when seeing a larger (or smaller) treatment effect, the researcher should not change her beliefs at all about whether the difference of means is explained more by the true treatment effect or by the bias in the research design. The magnitude of $\hat{\delta}$ conveys no information about whether or not it is driven by bias; instead, that belief depends entirely on how precise prior beliefs are about the treatment effect and the bias, as well as the correlation between the two. Two caveats apply to this claim. First, in section 6.3 we will show how different distributional assumptions about the priors may yield situations in which we learn something about whether the biased estimate is more or less driven by the true treatment effect depending on the size of the difference of means. Second, this discussion does not address the problems of publication bias or on so-called Type-M errors that come from underpowered statistical studies estimating small effects (Gelman and Carlin, 2014).

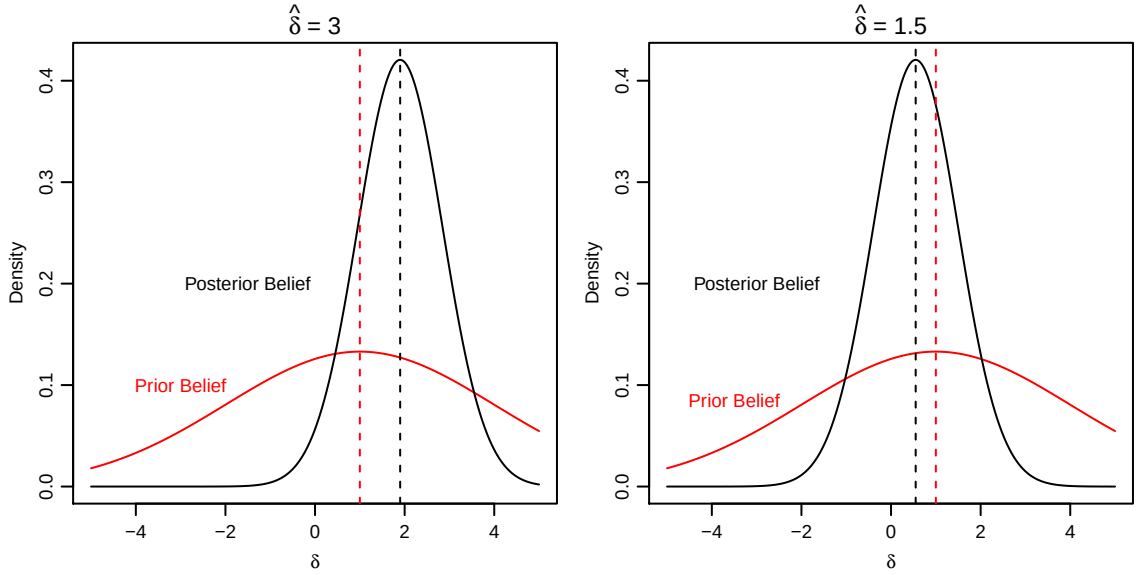
### 5.2 Revisiting The Motivating Example

We return now to our motivating example in which we seek to learn if the use of public transportation affects public support for government spending. Recall that we have estimate $\hat{\delta}$ that is the difference in mean support for government spending among public transportation uses and mean support for government spending among non-users. We would like to interpret this as an estimate of $\delta$, the effect of public transport use and support for government spending. But we suspect that this represents both the true effect of public transportation and some bias emerging from the fact that the people who support government spending also are the types who use public transportation. How shall we proceed to learn about $\delta$ from $\hat{\delta}$?

Our approach is to place priors beliefs on $\delta$ and calculate our posterior estimate of $\mu_\delta$ and $\overline{\sigma}^2_\delta$. Let us imagine a scenario in which we are relatively more uncertain about the treatment effect ($\sigma^2_\delta = 3$) than we are about the bias ($\sigma^2_\gamma = 1$), but our prior belief is that both are positive and uncorrelated with one another ($\mu_\delta = \mu_\gamma = 1, \rho = 0$). Observing a particular estimate of $\hat{\delta}$, what do we learn? We display the results visually in Figure 1 for two potential observed differences of means: $\hat{\delta} = 3$ and $1.5$. Where $\hat{\delta} = 3$, we have updated our estimate of the treatment effect upwards—and increased the precision of that estimate—based on a treatment effect that is substantially larger than our prior belief and the bias term. Where $\hat{\delta} = 1.5$, by contrast, our posterior mean belief about the treatment effect is smaller than the prior. This is because given the prior beliefs a treatment effect of $\hat{\delta} = 2$ is average, and lower values lead to lower beliefs about the treatment effect. However, $\hat{\delta} = 1.5$ is still a positive signal about the treatment effect since it is larger than the mean belief about bias.

In Figure 2 we show more generally how posterior beliefs about treatment and bias change given any particular observed estimate of $\hat{\delta}$. The left panel of Figure 2 shows how the beliefs about the treatment effect (thick line) and bias (thin line) change as a function of the observed difference of means for this example. When the observed difference of means is exactly $\mu_\gamma + \mu_\delta$ (here, 2), the mean of the posterior belief about the treatment effect is unchanged (as is the belief about bias).

Figure 1: Two Examples of Updating



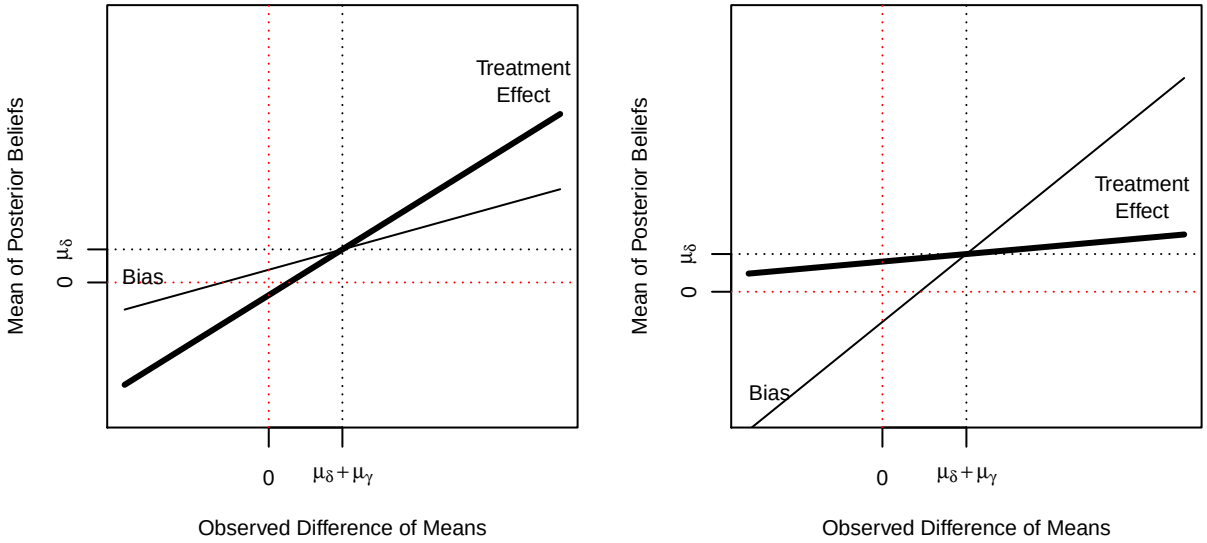

*Note*: Simulation values are $\sigma_\delta^2 = 3$, $\sigma_\gamma^2 = 1$, $\mu_\delta = \mu_\gamma = 1, \rho = 0$. Biased estimate $\hat{\delta} = 3$ (left) and $1.5$ (right).

Though, importantly, this is not a case where we "learn nothing", as the variance of the posterior belief goes down (i.e., there is second-moment learning).

For higher treatment effects, the mean of the belief about the treatment effect goes up, while for lower differences of means it goes down. The mean belief about bias moves in the same direction, but with a lower slope as there is less uncertainty about this parameter. The fact that both slopes are constant illustrates the fact that how the increase in beliefs gets "divided" between treatment and bias is constant. The right panel shows a similar picture but with the standard deviations flipped, so the researcher is now more uncertain about the bias at the outset. Now the bias update is steeper, indicating there is more learning about this parameter. Still, higher differences of means lead to a higher belief about the treatment effect.

An interesting contrast between these cases is what happens upon observing a "null result" with $\hat{\delta} = 0$. In the left panel of Figure 2, this leads to a much lower – in fact, negative – posterior mean belief about the treatment effect. This is because the researcher is relatively confident that there is a positive bias, and hence a zero difference of means points towards a negative treatment

Figure 2: Updating Beliefs about Treatment and Bias from the Observed Treatment Effect




*Note*: The thick line denotes mean posterior beliefs about the treatment effect, and the thin line denotes mean posterior beliefs about bias. In the left panel, $\sigma_\delta^2 = 3$, $\sigma_\gamma^2 = 1$, $\mu_\delta = \mu_\gamma = 1, \rho = 0$. In the right panel, the means and correlation are the same but $\sigma_\delta^2 = 1$ and $\sigma_\gamma^2 = 3$.

effect. On the other hand, in the right panel, the unexpectedly small zero difference of means leads to only a slightly lower belief about the treatment effect because this discrepancy from the prior is primarily attributed to bias.

Another important contrast between these examples is that we learn more about the treatment effect from the difference of means when we are more uncertain about the treatment effect than the bias. In particular, when $\rho = 0$, the ratio $\frac{\sigma_\delta^2}{\sigma_\delta^2+\sigma_\gamma^2}$ exactly determines the slope on the learning about the treatment effect and the degree to which the variance of the belief about the treatment effect shrinks.

# 6 Variants and Extensions

## 6.1 Sampling Error

The analysis with normal distributions can easily incorporate sampling error, by adding an error term so $\hat{\delta} = \delta + \gamma + \epsilon$. By an analogous calculation (assuming sampling error is mean zero and uncorrelated with the treatment effect and the bias in the prior) the posterior beliefs about the treatment effect and bias can still be expressed as written in Equations 3-4

$$m_\delta = \frac{\sigma_\delta^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2 + \sigma_\epsilon^2}$$
$$m_\gamma = \frac{\sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2 + \sigma_\epsilon^2}.$$

And the posterior belief about how much sampling error was in this estimate is $m_\epsilon(\hat{\delta} - \mu_\delta - \mu_\epsilon)$ where $m_\delta + m_\epsilon + m_\epsilon = 1$. So, the principle that the researcher must update positively on $\delta + \gamma + \epsilon$ as the treatment increases by one unit remains, though now some of the updating is soaked up by the error term as well.

## 6.2 Design Based on Subsamples

A frequent debate in empirical work is over how much we should focus on subpopulations where we can estimate treatment effects more precisely. In some research designs, we may be able to select a subsample of the observations where we know that the bias term is zero, even if our goal is to learn about the treatment effect for the entire sample. Perhaps a survey experiment is only feasible on a convenience sample; our prior belief is that there is no bias among those surveyed only. In the case of a regression discontinuity design, our prior belief is that there is minimal bias among the subsample of observations just above and below the threshold of the forcing variable (see Titiunik and Sekhon 2017 for a more complete discussion). In the case of instrumental variables, our prior belief is that there is minimal bias in the estimate of the average

treatment effect among the Compliers (those induced to take the treatment by the instrument). Other examples are possible.

Some existing approaches to generalizing from local average treatment effects to average treatment effects for some larger population include Aronow and Carnegie (2013) and Bisbee et al. (2017). These approaches rely on using observational data to characterize the similarity between the Compliers and the rest of the population. We proceed differently. In our terminology, suppose we have a subsample estimate:

$$\hat{\delta}^s = \delta^s + \gamma^s$$

If $\gamma^s$ is known to be small (or more precisely, our prior about it has a near-zero standard deviation) and there is no sampling error, then we can learn about $\delta^s$ with near certainty. This may be valuable information in and of itself, but if what we care about is the full-sample treatment effect $\delta$ then it is only valuable if we have a strong sense of the relationship between $\delta^s$ and $\delta$.

To know what the researcher learns about $\delta$ from $\hat{\delta}^s$, we need to know the prior belief about the subsample treatment effect and bias, and the relationship between these variables and the full sample treatment effect and bias. Let $\mu_{\delta^s}$ and $\mu_{\gamma^s}$ be the priors means of the subsample properties, with variances $\sigma^2_{\delta^s}$ and $\sigma^2_{\gamma^s}$. The update on $\delta$ conditional on $\hat{\delta}^s$ is normal with mean:

$$\overline{\mu}_\delta = \mu_\delta + m_{\delta^s}(\hat{\delta}^s - \mu^s_\delta - \mu^s_\gamma) \tag{7}$$

where

$$m_{\delta^s} = \frac{Cov(\delta, \hat{\delta}^s)}{Var(\hat{\delta}^s)} = \frac{\mathbb{E}[(\delta - \mu_\delta)(\delta^s - \mu_{\delta^s} + \gamma^s - \mu_{\gamma^s})]}{\mathbb{E}[(\delta^s - \mu_{\delta^s} + \gamma^s - \mu_{\gamma^s})(\delta^s - \mu_{\delta^s} + \gamma^s - \mu_{\gamma^s})]} = \frac{\sigma_\delta(\rho_{\delta,\delta^s}\sigma_{\delta^s} + \rho_{\delta,\gamma^s}\sigma_{\gamma^s})}{\sigma^2_{\delta^s} + 2\rho_{\delta^s,\gamma^s}\sigma_{\delta^s}\sigma_{\gamma^s} + \sigma^2_{\gamma^s}}$$

and the posterior variance of the estimate of $\delta$ is:

$$\sigma_\delta^2 - \frac{Cov(\hat{\delta}^s, \delta)^2}{Var(\hat{\delta})} = \sigma_\delta^2 - \frac{(\rho_{\delta,\delta^s}\sigma_\delta\sigma_{\delta^s} + \rho_{\delta,\gamma^s}\sigma_\delta\sigma_{\gamma^s})^2}{\sigma_{\delta^s}^2 + 2\rho_{\delta^s,\gamma^s}\sigma_{\delta^s}\sigma_{\gamma^s} + \sigma_{\gamma^s}^2} \tag{8}$$

where $\rho_{x,y}$ is the prior correlation between $x$ and $y$. (See the Appendix A for the full derivation.) So this estimate is preferred in the sense of doing a better job of shrinking the posterior variance of $\delta$ if and only if:

$$\frac{Cov(\hat{\delta}^s, \delta)^2}{Var(\hat{\delta}^s)} \geq \frac{Cov(\hat{\delta}, \delta)^2}{Var(\hat{\delta})}$$

An instructive special case is if there is no bias in the subsample, i.e., $\sigma_{\gamma^s} = 0$. If so:

**Proposition 3.** *If the subsample has no bias, then:*

$$m_{\delta^s} = \frac{\rho_{\delta,\delta^s}\sigma_\delta\sigma_{\delta^s}}{\sigma_{\delta^s}^2} = \frac{\rho_{\delta,\delta^s}\sigma_\delta}{\sigma_{\delta^s}}$$

*and*

$$\overline{\sigma}_\delta^2 = (1 - \rho_{\delta,\delta^s}^2)\sigma_\delta^2$$

*If, further, $\rho_{\delta,\gamma} = 0$, then the subsample is more informative about the treatment effect than the full sample if and only if:*

$$\rho_{\delta,\delta^s}^2 \geq \frac{\sigma_\delta^2}{\sigma_\delta^2 + \sigma_\gamma^2} \tag{9}$$

As long as the prior variances on $\gamma$ and $\delta$ are finite, the right-hand side of Equation 9 is strictly between 0 and 1, which has several consequences. First, and not surprisingly, if the sample treatment effect and the population treatment effect are completely independent ($\rho_{\delta,\delta^s}^2$), then the full

sample is always more informative.[4] On the other extreme, if the subsample treatment effect is perfectly correlated with the full sample ($\rho_{\delta,\delta^s} = 1$), then the subsample is always more informative than the full sample. In-between, there is always a critical threshold in this correlation such that if the treatment and subsample are sufficiently highly correlated, we learn more from the subsample than the full sample. This threshold is higher when the variance in the treatment effect is more driven by uncertainty about the (full sample) treatment effect than bias, i.e., when the full sample difference of means is primarily informative about the treatment effect.

### 6.3 Learning about Design Quality, Part 1

Many modern observational studies make a substantive or theoretical argument about why their key treatment variable is as-if random. Conversely, critiques of observational work frequently attack the plausibility of the as-if random claim. For example, Kocher and Monteiro (2016) argue that Ferwerda and Miller (2014)'s analysis of the effects of devolution on violence resistance in Vichy France, which relies on the as-if random placement of German double-track railroads, is subject to bias. Whatever the merits of this particular historical argument, we argue that from the perspective of estimating the causal effect of devolution, debates about whether the assumption of no selection at all is defensible are really disagreements about what prior belief we should hold about the bias term. However, the normal distribution assumptions here may not be appropriate, if we interpret such a debate as to whether the data generating process exhibits exactly zero bias, an event that receives zero probability even if the prior on the bias term is quite precise.

A natural generalization of the analysis above is to place a normal mixture distribution on the bias term. Here we analyze a simple case where the bias term is exactly zero with probability $\pi \in (0, 1)$, and is drawn from a normal distribution with mean $\mu_\gamma$ and variance $\sigma^2_\gamma$ otherwise.

[4] It may be hard to imagine a realistic case where there is zero correlation between the full sample and subsample correlation, as if nothing else the full sample correlation includes the subsample. Still, this could be reasonable as a limiting case where the subsample is very small and "distinct" from the full sample.

Formally, we can model this by writing the difference of means estimate as:

$$\hat{\delta} = \delta + \omega\gamma,$$

where $\omega = 1$ with prior probability $\pi$ ("biased estimate") and $\omega = 0$ with probability $1 - \pi$. If $\omega = 0$, then the learning about $\delta$ is simple as the second term drops out, so the true treatment effect is exactly the observed difference of means ($\delta = \hat{\delta}$). If $\omega = 1$, the learning about $\delta$ is the same as in the model analyzed in Section 5.

The novel feature of this new distributional assumption is that unlike the baseline case analyzed above, the researcher will learn about $\omega$ from $\hat{\delta}$. That is, we learn something about whether the difference of means is unbiased or not based on whether the observed $\hat{\delta}$ is relatively more or less likely to be generated by a biased or unbiased estimate. So, the average belief about the treatment effect as a function of $\hat{\delta}$ is:

$$\mathbb{E}[\delta|\hat{\delta}] = Pr(\omega = 0|\hat{\delta})\hat{\delta} + Pr(\omega = 1|\hat{\delta})(\mu_\delta + m_\delta(\hat{\delta} - \mu_\delta - \mu_\gamma)) \tag{10}$$

where $m_\delta$ is as defined in Section 5.

To derive the $Pr(\omega|\hat{\delta})$ terms, first observe the distribution of $\hat{\delta}$ under $\omega = 0$ is normal with mean $\mu_\delta$ and variance $\sigma_\delta^2$. The distribution of $\hat{\delta}$ under $\omega = 1$ is normal with mean $\mu_\delta$ and variance $\sigma_\delta^2 + \sigma_\gamma^2$. So, the posterior belief that the treatment effect is biased is:

$$Pr(\omega = 1|\hat{\delta}) = \frac{\pi \frac{1}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}} \phi\left(\frac{\hat{\delta}-(\mu_\delta+\mu_\gamma)}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}}\right)}{\pi \frac{1}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}} \phi\left(\frac{\hat{\delta}-(\mu_\delta+\mu_\gamma)}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}}\right) + (1-\pi)\frac{1}{\sigma_\gamma}\phi\left(\frac{\hat{\delta}-\mu_\delta)}{\sigma_\delta}\right)}, \tag{11}$$

and $Pr(\omega = 0|\hat{\delta}) = 1 - Pr(\omega = 1|\hat{\delta})$.

How do these beliefs change as a function of $\hat{\delta}$? If $\mu_\gamma = 0$, then the answer is straightforward, as both conditional distributions have mean $\mu_\delta$, but the variance is higher when the estimate is

potentially biased ($\omega = 1$). So, we should think the estimate is more likely to be unbiased when the difference of means is close to the prior mean, and more likely to be biased when observing a difference of means far from $\mu_\delta$. When $\mu_\gamma \neq 0$, then there is also a tendency to think the estimate is more likely to be unbiased when $\hat{\delta}$ is closer to $\mu_\delta$ and more likely to be biased when $\hat{\delta}$ is closer to $\mu_\delta + \mu_\gamma$. Combining these forces, there is always a critical value $\hat{\delta}^*$ where we should think the estimate is least likely to be biased, and the belief that the estimate is biased increases as the difference of means is further away from this value:

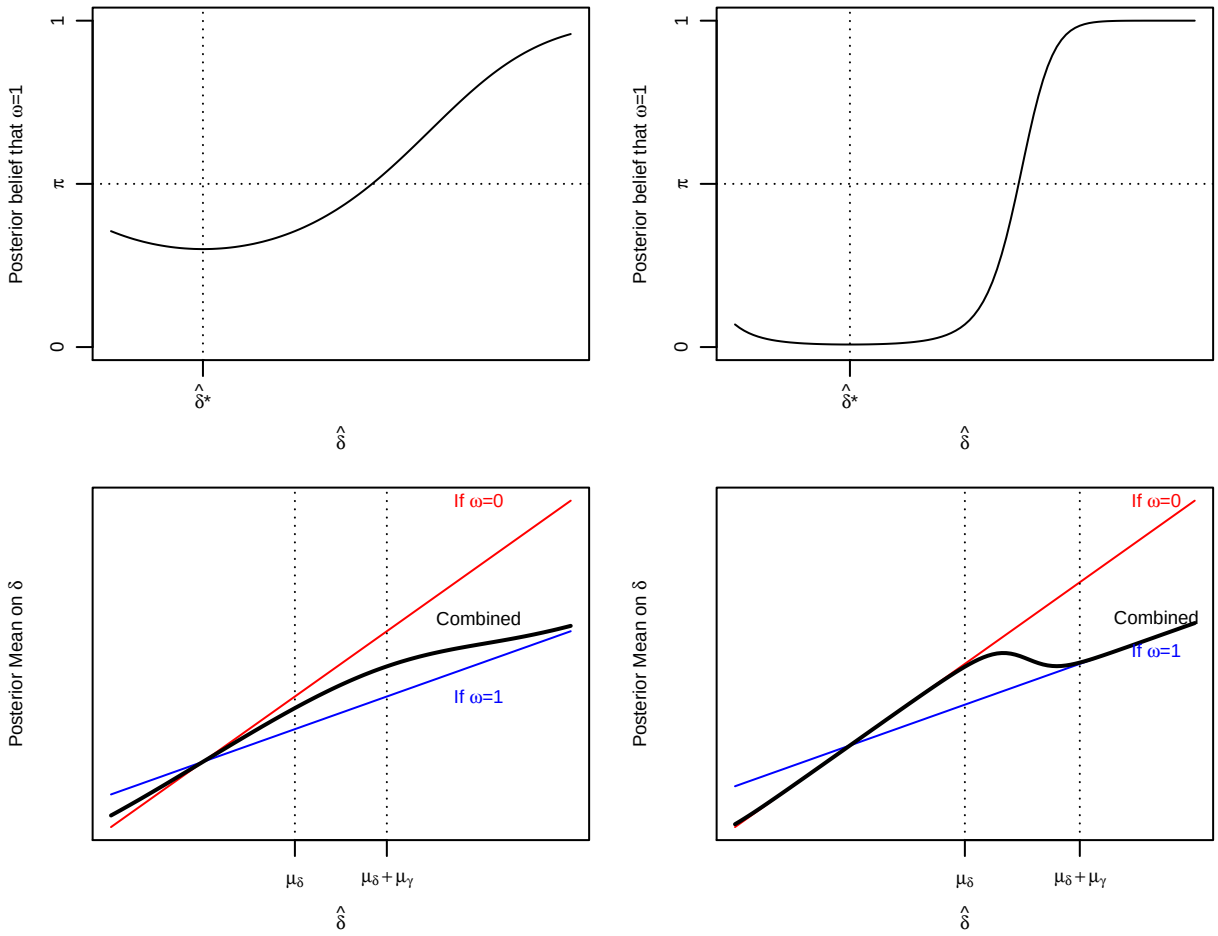
**Proposition 4.** *(i) There exists a* $\hat{\delta}^* \equiv \mu_\delta + \frac{\sigma_\gamma^2}{\sigma_{\hat{\delta}}^2}\mu_\gamma$ *such that* $Pr(\omega = 1|\hat{\delta})$ *is decreasing in* $\hat{\delta}$ *for* $\hat{\delta} < \hat{\delta}^*$*, and is increasing otherwise. (ii) As* $\hat{\delta} \to \pm\infty$*,* $Pr(\omega = 1|\hat{\delta}) \to 1$.

In this example, upon observing an "extreme" result, the researcher should be more likely to conclude that $\omega = 1$, meaning that there is bias. This also implies that beyond a certain point, when observing a higher difference of means, researchers should be more skeptical that the estimate is unbiased. And if a difference of means is associated with a more positive treatment effect with an unbiased estimate, the beliefs about the mean treatment effect can be locally *decreasing* in the observed difference of means.

Figure 3 shows examples where this does and does not happen. The top panels plot the posterior belief that the design is biased as a function of $\hat{\delta}$ for two simulations, and the bottom panels plot the mean of the posterior belief about $\delta$ as a function of $\hat{\delta}$. In these plots, the thin red line corresponds to the mean belief about $\delta$ conditional on $\omega = 0$, and the thin blue line conditional on $\omega = 1$. So, as shown in Equation 10, the overall belief is a weighted average of these two beliefs (thick black curve), where the weight is given by the value in the top panel. The difference between the left and right panels is that in the right panels, the prior belief about the bias term $\gamma$ is larger (if the design is in fact biased, $\omega = 1$).

In the bottom left panel, the thick black curve is always increasing, indicating that while higher treatments effects lead to more skepticism that the design is unbiased (over this range of difference of means), higher observed differences of means *also* lead to higher posterior beliefs about the

Figure 3: Learning about Design




*Note*: In all panels, $\sigma_\gamma = \sigma_\delta = 1$, $\pi = 1/2$, $\mu_\delta = 1$. In the left panels, $\mu_\gamma = 1$ and in the right panels $\mu_\gamma = 3$. In the bottom two panels, the red line denotes the mean posterior belief about $\delta$ when $\omega = 0$, and the blue line denotes the mean posterior belief about $\delta$ when $\omega = 1$. The thick black line denotes the *overall* posterior belief about $\delta$.

treatment effect. In the bottom right panel, the curve is decreasing for a range between $\mu_\delta$ and $\mu_\delta + \mu_\gamma$. This is because in this range, given larger prior beliefs about the magnitude of the bias, the increasing skepticism that the design is unbiased outweighs the direct effect of the difference of means being larger.

Returning to the disputed results on German double-track railways, if we interpret the goal of the original paper to be to estimate a causal effect of interest, then one might refocus this debate on the as-if random placement of German double-track railways away from the precise assumption of

random assignment, and instead about what might be learned about the causal effect of devolution on resistance given prior beliefs about the size effect and about how decisions to devolve respond to concerns about resistance. This would require a more thorough discussion of the estimated effect size in Ferwerda and Miller (2014) than that found in either that piece or the response, which could—under the framework outline in this section—lead researchers to refine their beliefs about the design itself as well as the size of the causal effects estimated from that design.

### 6.4 Learning about Design Quality, Part 2

The previous analysis show how learning should about treatment effects and bias when the bias term is drawn from a normal mixture. There are also scenarios where a normal mixture prior on the *treatment effect* makes sense.

Take a prominent recent example: Bem (2011), which purports to show the existence of ESP, a phenomenon most researchers doubt exists. The analog of our treatment effect $\delta$ varies across the nine experiments in this paper. A relatively straightforward one is a memory test for common words *followed by* a randomized treatment of practice on a subset of those words. So, the "psi" effect is estimated as a difference of means in recall of words which were later rehearsed versus not.[5] Here a skepticism of ESP would translate to a prior belief that this difference of means ($\delta$) is *exactly* zero.

The analysis of this case is similar to the previous section. Let the difference of means estimate be:

$$\hat{\delta} = \omega\delta + \gamma$$

where $\omega = 0$ means the difference of means is all bias (or sampling error), and when $\omega = 1$ the

[5] Another example asks participants to guess which of two windows contains a picture (initially concealed behind a curtain). In this experiment, the purported psi effect is the difference in the ability to correctly identify which contains the picture when it is of an erotic nature versus not.

estimate is the same as the main model. So, $\omega\delta$ is now the true treatment effect. Let $\pi = Pr(\omega = 1)$ be the prior belief that the treatment effect is non-zero.

The mean belief about the treatment effect given $\hat{\delta}$ can now be written

$$\mathbb{E}[\delta|\hat{\delta}] = Pr(\omega = 0|\hat{\delta})0 + Pr(\omega = 1|\hat{\delta})(\mu_\delta + m_\delta(\hat{\delta} - \mu_\delta - \mu_\gamma)) \tag{12}$$

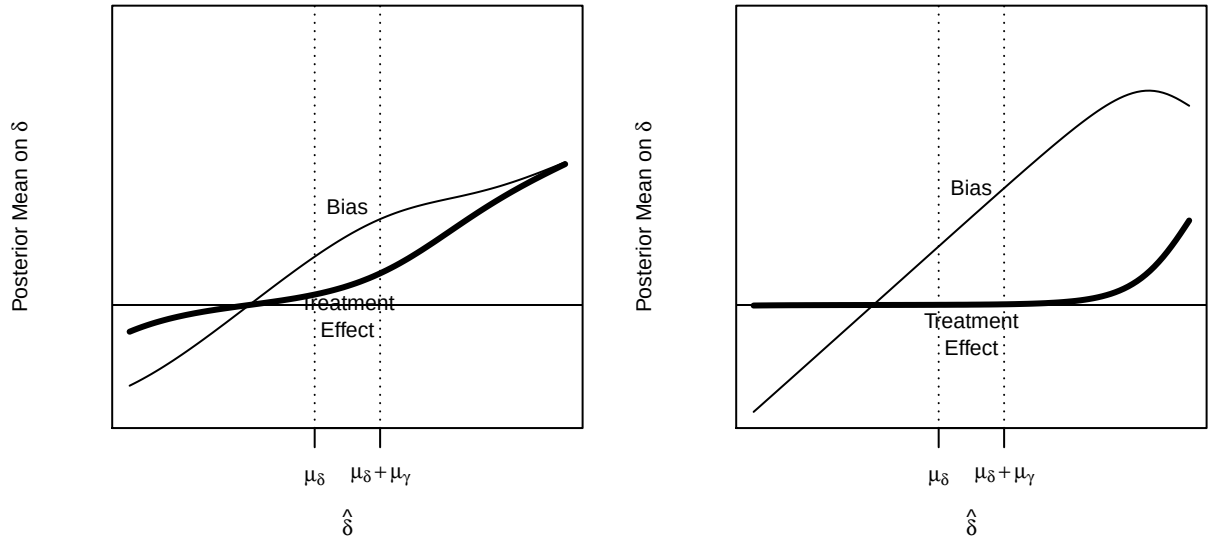
where $m_\delta$ is as defined in Section 5. The posterior belief that the treatment effect is non-zero is given by:

$$Pr(\omega = 1|\hat{\delta}) = \frac{\pi \frac{1}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}} \phi\left(\frac{\hat{\delta}-(\mu_\delta+\mu_\gamma)}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}}\right)}{\pi \frac{1}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}} \phi\left(\frac{\hat{\delta}-(\mu_\delta+\mu_\gamma)}{\sqrt{\sigma_\delta^2+\sigma_\gamma^2}}\right) + (1-\pi)\frac{1}{\sigma_\gamma}\phi\left(\frac{\hat{\delta}-\mu_\gamma)}{\sigma_\gamma}\right)}, \tag{13}$$

and $Pr(\omega = 0|\hat{\delta}) = 1 - Pr(\omega = 1|\hat{\delta})$.

Figure 4 illustrates two examples of this updating. Again, thick curves represent the mean of the posterior belief about $\delta$, and thin curves the mean of the belief about $\gamma$. In the left panel, the prior belief is that the treatment effect is exactly zero with probability .5, while in the right panel, that prior belief is that the treatment effect is exactly zero with probability .99. The curve is nonlinear (but monotone) in both panels, as different values of $\hat{\delta}$ change the belief about which part of the mixture $\delta$ is drawn from. In the left panel, this posterior belief changes the most for values somewhat above the expected $\hat{\delta}$ when $\delta \neq 0$, as in this range higher values of $\hat{\delta}$ push the researcher to believe that $\delta \neq 0$ and that the treatment effect is higher conditional on $\delta \neq 0$. In the right panel, where the researcher is highly skeptical that the treatment effect is non-zero (i.e., the "ESP case"), for the range of likely $\hat{\delta}$, any non-zero estimate of the treatment effect just allows the observe to learn about the bias (or error term). Only for extremely high values of $\hat{\delta}$ does the researcher start to believe that the treatment effect is truly different than zero.

Figure 4: Learning about Design, $\delta$ might be exactly zero




*Note*: The thick line denotes mean posterior beliefs about the treatment effect, and the thin line denotes mean posterior beliefs about bias. In both panels, $\sigma_\gamma = \sigma_\delta = \mu_\gamma = \mu_\delta = 1$. In the left panel $\pi = 1/2$, while in the right panel $\pi = 1/100$.

### 6.5 Why Might The Priors on Treatment and Bias be Correlated?

So far we have generally allowed the prior belief about the treatment effect and bias to be positively or negatively correlated, but have not provided intuitions for when either should be the case. In Appendix B we provide an explicit data generating process to explore this question. Unfortunately we find that there are cases where the prior belief on these variables should have a non-zero correlation, but in way which depends on several factors (e.g., whether there is positive nor negative selection into treatment, whether selection into treatment and baseline outcomes is positive or negative, and whether the treatment effect is *ex ante* likely to be positive or negative).

In our view the main lesson of this analysis is not that researchers should be able to clearly articulate what this prior correlation should be in the way we advocate they should be able to when considering the marginal distributions of $\delta$ and $\gamma$. However, in all of the simulations we have run, this prior correlation is rarely too extreme, so an assumption that it is equal to zero is unlikely to

be dangerous.[6]

## 7 Discussion and Conclusion

In this paper we have formulated the problem of learning about treatment effects from observational research as a standard problem of extracting information from a noisy signal. In contrast to existing approaches—which seek to eliminate the noise or somehow to bound it—we proceed in a Bayesian fashion by describing formally what it means to learn about a treatment effect and how prior beliefs about it and the bias allow us to update our posterior beliefs about that effect. Our analysis nests existing results such as Gerber, Green and Kaplan (2014)'s "illusion of learning from observation research" as special cases. It also allows us to describe what "learning" means in a more rigorous way, and to reveal some non-obvious ways that we can change our beliefs about treatment effects even with research designs that are known to be biased.

We conclude by proposing that this exercise is not merely a thought experiment or toy example constructed to illustrate what the problem of learning from observational research is, and how it might *in principle* be confronted. There are concrete and practical consequences to approaching learning about observational research as a task of extracting information from a noisy signal. We believe that these should affect current research practice. Some of these are easily implementable already; some of these require a shift in the ways that researchers think about design-based objections to observational results.

Begin with the way that researchers think about design-based objections. The current convention in seminars and referee reports is to identify possible threats to inference that may explain observational findings. If these possible objections are plausible, then the design is not credible and the findings suspect. We think that it is possible to move one step further by working through how much posterior estimates of the treatment effect of interest would change given on the ob-

[6]Recall this also implies the conditions where the estimate of $\delta$ is decreasing in $\hat{\delta}$ is unlikely to be met.

served result and prior beliefs. To do this, however, priors must be proper priors, which means that scholars must acknowledge that they *actually have* priors. We discussed in Section 4 why we think that researchers almost always have priors about treatment effects and bias terms, even if they are unwilling to specify them precisely. Here we make the implications plain: claiming "I believe that there is a confound that might explain this result" is equivalent to saying "here is a prior belief about that bias term." Equivalently, if you claim to have no prior belief about the bias term at all, then you could not object "I believe that bias in the research design likely explains this observed result."[7]

As in most areas of Bayesian analysis, many scholars may feel uncomfortable specifying their prior beliefs about treatment and bias terms. For sociological reasons, researchers might find themselves following conventions or rules of thumb. One convention that a community of scholars might adopt is a convention that the prior belief about the treatment effect is always a mean zero with a finite variance. One intuitive consequence of this convention is that the researcher's beliefs about the bias term determine which direction to update given any observed treatment effect ($\hat{\delta}$): the belief about the treatment effect will increase if and only if $\hat{\delta} > \mu_\gamma$

One non-intuitive consequence, however, is the precise degree of updating depends on the variances of $\delta$ and $\gamma$. To see why, start with the prior belief that $\delta = 0$. If the prior on $\gamma = 1$, then the update about the treatment effect is always positive just so long as $\hat{\delta} > 1$, but from equation 3 we know that just how large that update is will depend on $\sigma_\delta^2$ and $\sigma_\gamma^2$. It might be possible to accumulate information about the bias term in ways that allow researchers to have greater precision (less variance) in their beliefs about it. Our approach in this paper provides a straightforward way to interpret how precise our beliefs have to be to generate meaningful posterior updates about the treatment effect given any particular form of bias.

[7]It does not suffice to have a proper prior with a very large variance. A sincere objection to a result based on the possibility of bias is inconsistent with placing a large variance on the prior belief about that bias. Placing a large variance on the bias term is equivalent to saying that you do not hold very precise beliefs about it.

## Appendix A: Derivations and Proofs

**Derivations of Posterior Beliefs (section 5)** To derive the posterior beliefs about $\delta$ and $\gamma$, the joint distribution of these variables and $(\delta, \gamma, \hat{\delta})$ is a multivariate normal with mean vector $(\mu_\delta, \mu_\gamma, \mu_\delta + \mu_\gamma)$, and a covariance matrix:

$$\begin{array}{c} \\ \delta \\ \gamma \\ \hat{\delta} \end{array} \begin{array}{c} \begin{array}{ccc} \delta & \gamma & \hat{\delta} \end{array} \\ \begin{pmatrix} \sigma_\delta^2 & \rho\sigma_\delta\sigma_\gamma & \sigma_\delta^2 + \rho\sigma_\delta\sigma_\gamma \\ \rho\sigma_\delta\sigma_\gamma & \sigma_\gamma^2 & \sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma \\ \sigma_\delta^2 + \rho\sigma_\delta\sigma_\gamma & \sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma & \sigma_\delta^2 + \sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma \end{pmatrix} \end{array} \equiv \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

The right-hand side of this equation defines a partition of the covariance matrix, where $\Sigma_{22} = \sigma_\delta^2 + \sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma$.

The joint distribution of $(\theta, \delta)$ conditional on $\hat{\delta}$ is then jointly normal (Greene, 2008, p. 1014) with mean vector:

$$(\overline{\mu}_\delta, \overline{\mu}_\gamma) = = (\mu_\delta, \mu_\gamma) + \Sigma_{12}\Sigma_{22}^{-1}(\hat{\delta} - \mu_\delta - \mu_\gamma)$$

and covariance matrix:

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

which simplify to equations 3-5.

**Derivations of Posterior Beliefs (section 5)** The derivation of the posterior belief upon observing $\hat{\delta}^s$ follows the same logic as above, but now we need the joint distribution of $(\delta, \gamma, \hat{\delta}^s)$, which

is normal with mean $(\mu_\delta, \mu_\gamma, \mu_\delta^s + \mu_\gamma^s)$, and covariance matrix:

$$\begin{array}{c} \\ \delta \\ \gamma \\ \hat{\delta}^s \end{array} \begin{array}{c} \begin{array}{ccc} \delta & \gamma & \hat{\delta}^s \end{array} \\ \begin{pmatrix} \sigma_\delta^2 & \rho\sigma_\delta\sigma_\gamma & \rho_{\delta,\delta^s}\sigma_\delta\sigma_{\delta^s} + \rho_{\delta,\gamma^s}\sigma_\delta\sigma_{\gamma^s} \\ \rho\sigma_\delta\sigma_\gamma & \sigma_\gamma^2 & \rho_{\gamma,\delta^s}\sigma_\gamma\sigma_{\delta^s} + \rho_{\gamma,\gamma^s}\sigma_\gamma\sigma_\gamma^s \\ \rho_{\delta,\gamma^s}\sigma_\delta\sigma_{\gamma^s} & \rho_{\gamma,\delta^s}\sigma_\gamma\sigma_{\delta^s} + \rho_{\gamma,\gamma^s}\sigma_\gamma\sigma_\gamma^s & \sigma_{\delta^s}^2 + \sigma_{\gamma^s}^2 + \rho\sigma_{\delta^s}\sigma_{\gamma^s} \end{pmatrix} \end{array} \equiv \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

and so the joint distribution of $(\theta, \delta)$ conditional on $\hat{\delta}^s$ has mean vector

$$(\overline{\mu}_\delta, \overline{\mu}_\gamma) = = (\mu_\delta, \mu_\gamma) + \Sigma_{12}\Sigma_{22}^{-1}(\hat{\delta}^s - \mu_\delta^s - \mu_\gamma^s)$$

and covariance matrix:

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

which gives equations 7 and 8.

**Proof of proposition 4** For part (i), write the posterior belief as:

$$Pr(\omega = 1|\hat{\delta}) = \frac{1}{1 + \frac{Pr(\omega=0,\hat{\delta})}{Pr(\omega=1,\hat{\delta})}}.$$

This belief is increasing in $\hat{\delta}$ if and only if $\frac{Pr(\omega=1,\hat{\delta})}{Pr(\omega=0,\hat{\delta})}$ is increasing in $\hat{\delta}$. Since the log transformation is monotone, this is true if and only if:

$$\begin{aligned} &\log Pr(\omega = 1, \hat{\delta}) - \log Pr(\omega = 0, \hat{\delta}) \\ &= k + \left(\frac{\hat{\delta} - \mu_\delta - \mu_\gamma}{\sqrt{\sigma_\delta^2 + \sigma_\gamma^2}}\right)^2 - \left(\frac{\hat{\delta} - \mu_\delta}{\sigma_\delta}\right)^2 \end{aligned} \tag{14}$$

is increasing in $\hat{\delta}$ (where $k$ is a constant). Differentiating and rearranging gives this is true when:

$$\begin{aligned}
\frac{\hat{\delta} - \mu_\delta - \mu_\gamma}{\sigma_\delta^2 + \sigma_\gamma^2} - \frac{\hat{\delta} - \mu_\delta}{\sigma_\delta^2} &\geq 0 \\
\sigma_\delta^2(\hat{\delta} - \mu_\delta - \mu_\gamma) &\geq (\sigma_\delta^2 + \sigma_\gamma^2)(\hat{\delta} - \mu_\delta) \\
\sigma_\gamma^2\mu_\delta - \sigma_\delta^2\mu_\gamma &\geq \sigma_\gamma^2\hat{\delta} \\
\hat{\delta} &\leq \frac{\sigma_\gamma^2\mu_\delta - \sigma_\delta^2\mu_\gamma}{\sigma_\gamma^2}
\end{aligned}$$

For part (ii) note that as $\hat{\delta} \to \pm\infty$, (14) approaches $\infty$, and so $\frac{Pr(\omega=0,\hat{\delta})}{Pr(\omega=1,\hat{\delta})} \to 0$ and $Pr(\omega = 1|\hat{\delta}) \to 1$

## Appendix B: Analysis of the Prior Correlation $\rho$.

Here is a data generating process which aims to provide insight into when the prior belief about the correlation between the average treatment effect on the treated ($\delta$) and the bias in the naive estimate with respect to this parameter ($\gamma$) should be positive or negative.

A challenge of this exercise is that – for natural ways of writing down the data generating process – both the ATET and bias are endogenous outcomes. Further, generically speaking there are multiple parameters which affect $\delta$ and $\gamma$, and the observed correlation between these across cases may depend on which exogenous parameter drives the changes is these variables.

To limit scope, we focus on the case where the "core" parameter of uncertainty is the average treatment effect (not just on the treated); if we were not uncertain about this, as if we were uncertain about this the who exercise would be pointless. That is, we will ask "as the average treatment effect increases, how do $\delta$ and $\gamma$ change?".

We also focus on the case where the treatment effect is equal across units: $\delta_i = \delta$. So the ATET and ATE are the same, but there may be bias in the naive estimator due to correlation between baseline outcomes and treatment. Some analysis of the case where the treatment effect is

heterogenous across units indicates it leads to similar conclusions as this simpler case.

The question of whether we should expect a prior correlation between $\delta$ and $\gamma$ can be answered by checking how the bias in the naive estimator changes as the treatment effect changes. If $\gamma$ is increasing in $\delta$ we should expect them to be positively correlated, if $\gamma$ is decreasing in $\delta$ we should expect a negative correlation.

Let the measured outcome for unit $i$ as a function of the treatment status be:

$$y_i(t_i) = y_0 + \delta t_i + \nu_i, \tag{15}$$

where $\nu_i$ is a mean zero random variable, so $y_0$ is the average baseline outcome (if $t_i = 0$) and $\nu_i$ is the individual deviation from this average.

Suppose whoever decides the treatment status has a utility function (the "decision-maker", or "DM") of the form:

$$u(t_i, y_i) = \alpha y_i + (\beta + \eta_i) t_i \tag{16}$$

So, when $\alpha > 0$ the DM likes outcomes to be higher, and when $\alpha < 0$ the DM likes outcomes to be lower. There is also an intrinsic preference associated with the treatment $\beta + \eta_i$. Assume $\eta_i$ is mean zero, so $\beta$ represents whether the treatment is generally good or bad, and $\eta_i$ the individual specific deviation from the average.

Unit $i$ will get treated if $u_i(1, y_i(1)) > u_i(0, y_i(0))$, or:

$$\alpha(y_0 + \delta + \nu_i) + \beta + \eta_i > \alpha(y_0 + \nu_i) \tag{17}$$

$$\eta_i > -(\alpha\delta + \beta) \tag{18}$$

So, the bias in the naive estimator can be written:

$$\gamma = \mathbb{E}[y_i(0)|t_i = 1] - \mathbb{E}[y_i(0)|t_i = 0] \tag{19}$$

$$= \mathbb{E}[\nu_i|\eta_i > -(\alpha\delta + \beta)] - \mathbb{E}[\nu_i|\eta_i < -(\alpha\delta + \beta)] \tag{20}$$

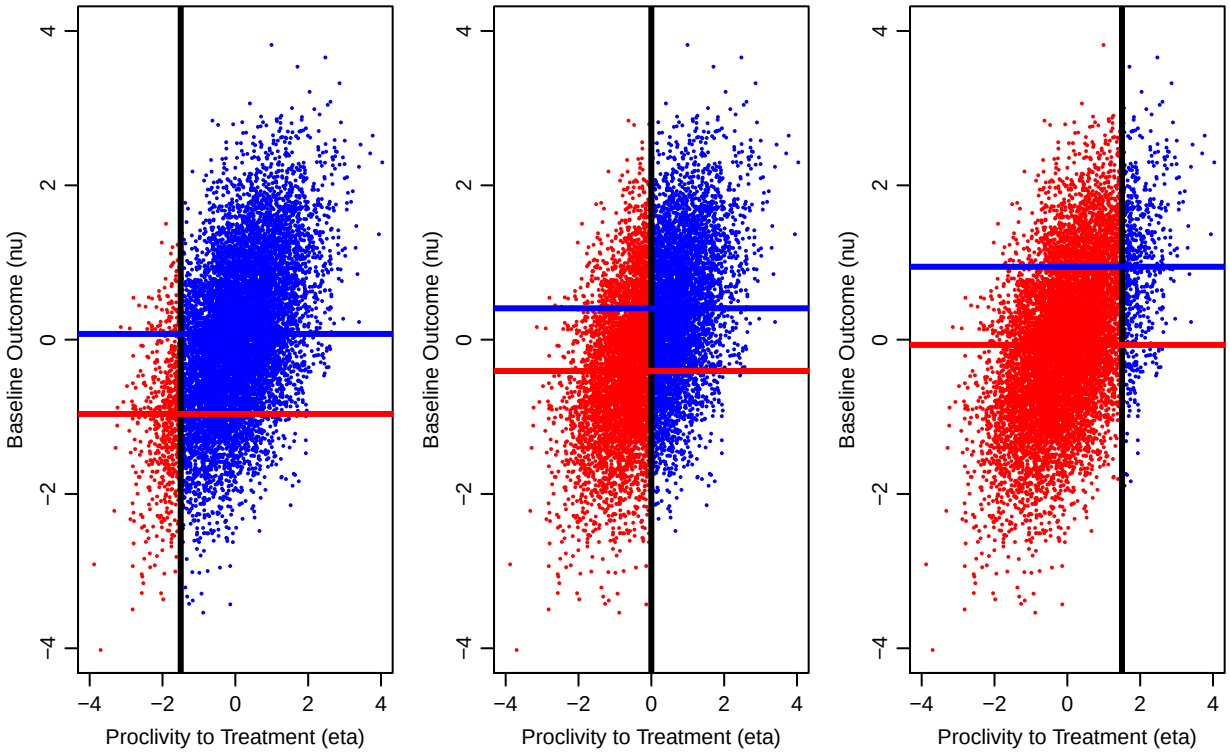
If $\eta_i$ and $\nu_i$ are uncorrelated (i.e., there is no correlation between the baseline outcome and proclivity to treatment), then this term will always be 0.

If $\eta_i$ and $\nu_i$ are positively correlated, then $\mathbb{E}[\nu_i|\eta_i > -(\alpha\delta + \beta)] - \mathbb{E}[\nu_i|\eta_i < -(\alpha\delta + \beta)]$ will generally be positive, since the first term selects on cases where $\eta_i$ is high (and hence $\nu_i$ will tend to be high). Conversely, if $\eta_i$ and $\nu_i$ are negatively correlated, then the bias will tend to be negative.

However, the question we are concerned with is how $\delta$ affects $\gamma$. The effect here is subtle. First, note that changes in $\delta$ matter in how they affect the threshold in $\eta_i$ which selects into treatment, call this $\tau = -(\alpha\delta + \beta)$. If $\alpha > 0$, this threshold is decreasing in $\delta$, and if $\alpha < 0$ the threshold is increasing in $\delta$. If $\alpha = 0$, the naive estimator will be biased, but the magnitude of the bias is not a function of $\delta$ and so we should not expect $\delta$ and $\gamma$ to be correlated.

How does changing the threshold of $\eta_i$ which determines treatment status affect the bias term? To streamline, focus on the case where $\nu_i$ and $\eta_i$ are positively correlated. More specifically, suppose they are drawn from a multivariate normal distribution with correlation $\rho > 0$. If so, *both* $\mathbb{E}[\nu_i|\eta_i > \tau]$ and $\mathbb{E}[\nu_i|\eta_i < \tau]$ are increasing in $\tau$.

The figure below illustrates. All panels plot draws of $\nu$ and $\eta$, which are positively correlated. The difference between the panels is the threshold in $\eta_i$ at which a unit gets treated. As the threshold increases, both groups on average have higher baseline outcomes (a Simpson's Paradox type effect).




For the normal case, this difference is smallest when $\tau = 0$, and is decreasing up to this point and increasing after.

Combining, this simulation suggests that when $\rho > 0$, then $\gamma$ is increasing in $\tau$ if and only if $\tau < 0$, and $\tau$ is increasing in $\delta$ if and only if $\alpha < 0$. We suspect analytic results along these lines are possible, but the main message is "it's complicated".

## References

Altonji, Joseph G., Todd E. Elder and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1):151–184.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical

Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2):3–30.

Aronow, Peter M. and Allison Carnegie. 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable." *Political Analysis* 21(4):492–506.

Bem, Daryl J. 2011. "Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of personality and social psychology* 100(3):407.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches and Cyrus Samii. 2017. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect." *Journal of Labor Economics* 35(S1):S99–S147.

DellaVigna, Stefano and Devin Pope. 2016. Predicting experimental results: who knows what? Technical report National Bureau of Economic Research.

Ferwerda, Jeremy and Nicholas L. Miller. 2014. "Political Devolution and Resistance to Foreign Rule: A Natural Experiment." *American Political Science Review* 108(3):642–660.

Gelman, Andrew and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9(6):641–651.

Gerber, Alan S, Donald P Green and Edward H Kaplan. 2014. "The illusion of learning from observational research." *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* pp. 9–32.

Gill, Jeff and Lee D Walker. 2005. "Elicited priors for Bayesian model specifications in political science research." *The Journal of Politics* 67(3):841–872.

Greene, William H. 2008. *Econometric Analysis, Sixth Edition*. Prentice Hall.

Kocher, Matthew A. and Nuno P. Monteiro. 2016. "Lines of Demarcation: Causation, Design-Based Inference, and Historical Research." *Perspectives on Politics* 14(4):952–975.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Oster, Emily. 2017. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* .

Rosenbaum, Paul R. 2002. *Observational Studies*. Second ed. New York: Springer.

Titiunik, Rocio and Jasjeet S. Sekhon. 2017. On Interpreting the Regression Discontinuity Design as a Local Experiment. In *Regression Discontinuity Designs: Theory and Applications*, ed. M.D. Cattaneo and J. C. Escanciano. Emerald Publishing Limited pp. 1–28.