Learning from Biased Research Designs

Andrew T. Little, University of California, Berkeley
Thomas B. Pepinsky, Cornell University

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. Our core contention is that any sincere claim that a research design contains bias entails a belief about what that bias is. Once this belief is specified (along with a prior belief about the causal effect), what one should learn from a potentially biased estimate can be derived from Bayes's 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 several methodological debates and practical problems confronting applied researchers.

The credibility revolution (see Angrist and Pischke 2010) has transformed empirical research in the social sciences. Whereas a previous generation's quantitative research established correlations among systems of variables and then inferred causal interpretations to those results, progress in statistical methods for causal inference has encouraged researchers to replace "kitchen sink" or "garbage can" regressions with statistical models designed to provide unbiased estimates of well-defined causal quantities (see Samii 2016). The challenge of this revolution is that it relies on strong research designs that require assumptions about the data-generating process for estimates to have their desired causal interpretation. In practice, debates about what we learn from empirical research are debates about those assumptions, or about what we *can* learn about from those designs.

Such debates tend to have an all-or-nothing flavor, with the most committed revolutionaries holding that if we are uncertain about the assumptions of the research design, then we learn nothing from the findings. But if ignoring the assumptions for identification is like sticking one's head in the sand, then refusing to learn anything from estimates without near-perfect identification is akin to staring at the sun and then complaining about not being able to see.

In this paper, we outline a different perspective, rooted in a Bayesian framework. Most empirical work in contemporary political science involves trying to learn about a treatment effect or causal effect δ. This is frequently done by calculating a difference of means or estimating a regression that returns an estimate $\hat{\delta}$, which is the parameter of interest plus some amount of bias, denoted γ: $\hat{\delta} = \delta + \gamma$. If the research design is "credible," meaning that it has no bias, then $\gamma = 0$ and we can learn that $\hat{\delta} = \delta$. Arguments about research design and the "credibility" of estimates of causal effects frequently center around whether assumptions about the bias term γ 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"). The challenge is, what we can learn when the credibility of a research design is unclear?

We draw a parallel between the problem of inferring δ from plausibly biased research designs and of learning from biased signals commonly employed in formal theory. Prominent formal theories about topics such as expert advice (Calvert 1985), accountability (Fearon 1999), and protest (Bueno De Mesquita 2010) share an analogous structure. An individual cares about some fact about the world, represented with a random variable δ. She starts with a prior belief about δ and then observes a signal, call this $\hat{\delta}$, which is contaminated by noise, represented with another random variable γ. A tractable way to formulate the problem is that $\hat{\delta} = \delta + \gamma$. So, a decision maker observing expert advice or a voter observing a signal of incumbent performance seeks to extract information about δ by observing $\hat{\delta} = \delta + \gamma$. The same logic that guides our decision maker in inferring incumbent performance from a plausibly biased signal, we argue, can guide

Andrew T. Little (andrew.little@berkeley.edu) is assistant professor of political science at the University of California, Berkeley, Berkeley, CA 94720. Thomas B. Pepinsky (pepinsky@cornell.edu) is Walter LaFeber Professor at Cornell University, Ithaca, NY 14853.

An online appendix with supplementary material is available at https://doi.org/10.1086/710088.

The Journal of Politics, volume 83, number 2. Published online April 19, 2021. https://doi.org/10.1086/710088

researchers who seek to infer causal effects from plausibly biased research designs.

The primary goals of this paper are expository and pedagogical. We present a framework for reasoning through imperfect research designs that clarifies how different approaches to causal inference will confront the problem of learning from observational data and imperfect research designs. Our core contention is that any sincere claim that a research design is biased requires a belief about what that bias is. Once we specify that belief (and a prior belief about the causal effect of interest), the question of how to learn from imperfect research maps exactly onto the model of learning from a biased signal. We find that in a basic setup this learning is at least monotonic, in the sense that higher reported estimates increase beliefs about causal effects. However, we also highlight some scenarios—in particular, when results may be driven by mistakes or even fraud, or when observers are uncertain about what specifications the researcher ran but did not report—where this monotonicity can break down.

We begin with a motivating example that is common in quantitative political science research: a causal variable of interest in an observational study that is likely confounded by unobserved heterogeneity. With this example in hand, we outline various contemporary strategies for inferring causal effects given the possibility of selection or omitted variable bias. We then introduce a formal model of the learning process that shows—given researcher beliefs—how we might use Bayesian reasoning to infer the causal effects in such designs. After working through this stylized model, we show how our approach can shed light on more practical and complex issues common to applied empirical work. In doing so, we bring a common structure to thinking about questions like how publication bias affects what we should learn from studies, when it is desirable to learn from subsamples or experiments with external validity concerns, and how the potential for major research malpractice (purposeful or not) affects how we should interpret the results that we see. In the concluding section, we draw out the implications of this argument for empirical researchers.

A MOTIVATING EXAMPLE

Consider a question that has received considerable attention recently: whether adopting limited democratic institutions causes authoritarian countries to have different political and economic outcomes (e.g., Gandhi 2008; Pepinsky 2014; Rivera 2016). To fix ideas, suppose a researcher is investigating whether holding elections causes changes in repression. This is a key question in the emerging literature on authoritarianism, and if it could be shown that holding elections causes authoritarian regimes to perpetrate less violence against their citizens, then this would be an argument in favor of supporting electoral institutions even when they do not generate truly democratic competition. It is hard to randomly assign electoral institutions to authoritarian regimes, and the researcher suspects that, say, asking undergraduates to take on the role of a dictator or some other experiment is too unrealistic and conceptually distant from the problem at hand to yield insights about repression. So, she proceeds with observational data, say, with variables measuring physical integrity rights as well as a binary variable D that captures whether each regime holds elections or not. She then runs a standard regression of $Y = \delta D + \epsilon$.

If we assume $\text{Cov}(D, \epsilon) = 0$ and ϵ is mean zero, then the estimate $\hat{\delta}$ from that regression is an unbiased estimate of the effect of authoritarian elections on mass repression. However, the former is an implausible assumption. The decision to hold elections is a strategic choice (e.g., Little 2017), and there are good reasons to suspect that the types of regimes that hold elections are also likely to be the types of regimes that face stronger constraints against—or have fewer reasons for—murdering their citizens (see Pepinsky 2014). If so, the estimate $\hat{\delta}$ does not represent the causal effect of authoritarian elections. It represents the sum of that effect and the effect that can be attributed to nonrandom selection. This matters for the researcher because she would not advocate for introducing elections to nonelectoral dictatorships if both authoritarian institutions and repression are the products of more fundamental sociopolitical dynamics.

Call a latent variable that represents that selection process U. If the researcher could observe U—all the factors that jointly affect both whether regimes hold elections and their proclivity for repression—she would condition on them in a regression of $Y = \delta D + \beta U + \epsilon$. She knows, however, that by a standard calculation of omitted variable bias she can represent the naive estimate as $\hat{\delta} = \delta + \beta[\text{Cov}(D, U)/\text{Var}(D)] + \epsilon$; that is, as the sum of the true effect of institutions plus the effect attributable to selection weighted by the relationship between selection and treatment and the sampling error. Relabeling $\beta[\text{Cov}(D, U)/\text{Var}(D)] = \gamma$, if either incentives to repress are unrelated to institutions ($\text{Cov}(D, U) = 0$) or incentives to repress actually have no relationship with repression ($\beta = 0$), then $\gamma = 0$ and $\hat{\delta} = \delta + \epsilon$. Under these assumptions, $\hat{\delta}$ is an unbiased estimate of δ.

LITERATURE

How might researchers proceed to learn about the effect of authoritarian institutions on repression if we suspect that $\gamma \neq 0$? The dominant approach in contemporary political science and applied microeconomics focuses on design. The

general aim 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 a subpopulation of authoritarian regime where institutions were assigned by a known process that is unrelated to incentives to repress, or where a surprise event removes elections without changing incentives to repress, or a related strategy. The result is an estimate of a local average treatment effect (LATE), which is an unbiased estimate of δ 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 the section "When Should We Learn More from Designs Based on Subsamples?"

Another approach is to identify an interval or range of values in which δ may lie based on minimal assumptions about unobserved features of the data (Manski 1995). This "partial identification" approach can be valuable when it yields results that are substantively useful even if the precise value of δ remains unknown; for example, one may discover that the range of δ is strictly positive. Common sources of leverage that can place bounds on the identification region are prior information about the logically possible range of δ, the distributional characteristics of the variables themselves (Manski 1990), or shape restrictions on the treatment response function (Manski 1997). Partial identification approaches are powerful, but by imposing complete agnosticism about γ, they omit important information that researchers can use to sharpen their inferences.

A third approach examines the sensitivity of $\hat{\delta}$ to various assumptions about γ. 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 probabilities yield small changes in $\hat{\delta}$, then one may use $\hat{\delta}$ to learn about δ even without precise information about γ. 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 γ on $\hat{\delta}$ without any further information about γ.

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

The closest methodological literature to our approach is that on Bayesian inference with partially identified models; see Gustafson (2015) for a recent treatment and McCandless, Gustafson, and Levy (2007) for an application to unobserved confounders in regression. Broadly, we apply this idea to a difference of means test and other designs common in contemporary political science, and 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. The most related precursor to this paper from the social sciences 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" attitude described previously. We discuss the areas of overlap in our results and theirs in the section "A Model of Learning with Normal Distributions," but here we highlight several differences. First, from a modeling perspective, we allow for more general prior beliefs, both by allowing for correlation between the bias and treatment priors when using normal distributions and by presenting results without the assumption of normality. Second, and more substantively, we focus attention on cases where researchers' 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 several prominent debates about research design not addressed in Gerber et al. (2014).

A MODEL OF LEARNING FROM BIASED DESIGNS

The key to our approach is to represent skepticism about the credibility of a research design as a prior belief that the research design is biased to some degree. A better way to describe our researcher's concern about identifying the causal effect of authoritarian institutions is that her prior belief about γ—the bias term—places high probability on values meaningfully far from zero. If her main concern is that regimes with elections tend to score better in measures of physical integrity rights for noncausal reasons, then her prior beliefs should place more probability on positive values of γ. Once we specify a prior belief about the joint distribution of δ and γ, it is straightforward in principle to use Bayes's rule to determine what our posterior beliefs about the true causal effect δ change, given those prior beliefs, if we observe a particular estimate $\hat{\delta}$.

Because specifying prior beliefs is essential to our argument, we first consider the objection that it is too hard or too unnatural to specify informative prior beliefs about parameters that we wish to learn about. Most scholars are uncomfortable specifying their prior beliefs. But 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.

Yes, you have priors

In practice—in the sense of "what people tend to do when estimating statistical models"—empirical researchers using Bayesian techniques usually employ "noninformative" (typically, high variance) and "neutral" (typically, mean 0) priors. However, the way scholars actually discuss research reveals that they do have informative (and often nonneutral) 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. Imagine, for example, that there were an entirely credible research design to estimate the effect of authoritarian institutions on repression. In this context, to have no prior belief about the effect of institutions is to say that no result—hugely positive effects of elections, hugely negative effects of elections, or no effect at all—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. Researchers have prior beliefs about treatment effects.

What about prior beliefs about bias? One can represent complete ignorance about bias by examining the limiting case as the prior approaches an improper uniform prior (which is what many results in Gerber et al. [2014] do). However, the way that scholars debate how to interpret empirical studies indicates that they do have beliefs about the direction and magnitude of bias. To say "I'm worried that your estimate is biased because of (insert confound/sampling/design issue)" is to say "I have a prior belief that γ 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 inconsistent with how scholars actually debate the merits of different research designs.[2]

Still, one may be uncomfortable translating these prior beliefs into a specific prior distribution (Gill and Walker 2005). And with good reason: in many cases there are multiple potential sources of bias, thinking through their magnitude is not easy, and there is certainly no guarantee that well-intentioned researchers would all agree on the appropriate prior belief. Those who have read different literatures, or find different theories more plausible, can certainly hold different prior beliefs about treatment effects or biases of research designs and will typically hold different posterior beliefs. More cynically, one can view the common response to identification concerns of claiming "yes, but the most likely forms of bias would go against our result" as equivalent to arguing "the correct prior belief about γ has the opposite sign as the reported treatment effect, so your posterior belief about the treatment effect should be even higher." Our analysis will show that this inference is warranted if one accepts such a prior belief.

This example of disagreement about prior beliefs raises the possibility of selective reporting (or emphasis) of potential biases. A motivated researcher might selectively underreport prior beliefs about bias in ways that are consistent with his or her preferred beliefs. Avoiding this kind of challenging and subjective process is why credible research designs are so powerful. And we do not propose that we can "fix" or "solve" bias simply by thinking about our prior beliefs about it. Rather, we propose that in domains where credible estimates are elusive, coming up with the best research designs possible and then thinking about how to update our beliefs from them may be the best we can do. 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. Even if

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

2. See Dunning (2008) for a discussion of how to place natural experiments on a "continuum of credibility."

researchers do not agree on their prior beliefs, they should agree on how to combine these priors with the results of studies to learn.

Once we acknowledge that researchers truly do have prior beliefs about effect size and bias, we can show how to use them to learn from imperfect designs. To do so, we start with a simple model of learning that begins with prior beliefs about δ and γ.

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). To set aside issues of sampling error (a point revisited in the section "What If There Is Sampling Error or Multiple Estimates?"), suppose a researcher observes the realized outcomes for the entire population. She then computes the standard difference of means estimate $\hat{\delta} = \mathbb{E}[Y_1^1|D = 1] - E[Y_0^0|D = 0]$ (subscripts referring to actual treatment status). By a standard calculation, this 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.

We focus on this kind of estimate since the relationship between the treatment effect and bias is particularly tidy and easy to interpret. However, it bears emphasis that we can always write an estimate of a causal effect as equal to the truth plus a bias term. This is simply an accounting identity, though of course depending on the estimand and estimate in question the bias term will take on different meaning. However, in the section "What If the Study Might Be Completely Flawed?" we discuss some cases where it makes more sense to think of the estimate as potentially unrelated to the treatment effect.

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

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

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 $f_\delta(\delta)$. Unless the difference of means is independent of the true treatment effect, there is always a difference between prior and posterior beliefs—and, thus, there is 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 δ changes as a function of $\hat{\delta}$.
2. To measure "second-moment learning," we study how much the variance of the posterior belief about δ 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

$$\text{Var}[\delta] - \mathbb{E}[\text{Var}[\delta|\hat{\delta}]] = \text{Var}[\mathbb{E}[\delta|\hat{\delta}]]. \quad (1)$$

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

The unit principle

Finally, before placing any distributional assumptions on the prior, we emphasize a simple identity about how beliefs about treatment and bias should jointly change as estimates change. 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, $\partial\mathbb{E}[\delta|\hat{\delta}]/\partial\hat{\delta} = 0$.

More generally, a question we might ask is, How do our beliefs about the treatment effect and bias change when the difference of means increases? In terms of how the mean of our beliefs about these variables changes, a property of this answer is immediate from the linearity of expectations. In particular, if $\hat{\delta} = \delta + \gamma$, and both δ and γ have a finite expectation, then

$$\frac{\partial\mathbb{E}[\delta|\hat{\delta}]}{\partial\hat{\delta}} + \frac{\partial\mathbb{E}[\gamma|\hat{\delta}]}{\partial\hat{\delta}} = \frac{\partial\mathbb{E}[\delta + \gamma|\hat{\delta}]}{\partial\hat{\delta}} = \frac{\partial\mathbb{E}[\hat{\delta}|\hat{\delta}]}{\partial\hat{\delta}} = 1. \quad (2)$$

Working backward in (2), 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 between 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 ($\partial\mathbb{E}[\delta|\hat{\delta}]/\partial\hat{\delta} = 1$ and $\partial\mathbb{E}[\gamma|\hat{\delta}]/\partial\hat{\delta} = 0$, or a "credible estimate") or only on the bias ($\partial\mathbb{E}[\delta|\hat{\delta}]/\partial\hat{\delta} = 0$ and $\partial\mathbb{E}[\gamma|\hat{\delta}]/\partial\hat{\delta} = 1$, 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 this unit principle is inviolable.

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 δ (the treatment effect) and γ (bias). A natural place to start is to assume that these two random variables are drawn from a multivariate normal distribution. We discuss some scenarios when this is likely not a good approximation and consider several alternative distributional assumptions in the section "What If the Study Might Be Completely Flawed?"

Consider the case with no sampling error and let the prior distribution on (δ, γ) be a multivariate normal with mean vector (μ_δ, μ_γ). The prior variances of the individual variables are σ_δ^2 and σ_γ^2, and the covariance is $\rho\sigma_\delta\sigma_\gamma$, where ρ is the correlation between the prior belief about bias and treatment. We provide some intuitions for when this correlation might be nonzero in the section "What If There Is Sampling Error or Multiple Estimates?"

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 δ and γ. Upon observing a smaller difference of means, she will generally update negatively on both.

First-moment learning. By a standard property of multivariate normal distributions, the posterior belief about (δ, γ) upon observing $\hat{\delta}$ is normally distributed with means

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

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

where

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

$$m_\gamma = \frac{\mathrm{Cov}(\gamma, \hat{\delta})}{\mathrm{Var}(\hat{\delta})} = \frac{\sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2}.$$

(See app. A.1 for the full derivation; appendix is available online.) 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 on the point estimate on that variable. When $m_\delta > 0$—that is, when the covariance between δ and $\hat{\delta}$ is positive—the researcher has a more positive belief about the true treatment effect upon observing a higher difference of means.

We now relate those theoretical results to some basic intuitions about how we learn from new data. First, consider some limiting cases. On one extreme, when we already know the treatment effect δ with certainty ($\sigma_\delta \to 0$), we put no weight on the results from a new study ($m_\delta \to 0$).

At the other extreme, if we are certain about the magnitude of the bias parameter γ for a particular research design, we can then 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 $\bar{\sigma}_\delta^2 \to 0$. Thus, the right-hand side of 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 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 $\hat{\delta}$.

There is a particularly simple way to express how much we should update our beliefs abut δ when $\rho = 0$, that is, the prior distributions of δ and γ are independent of one another. If so $m_\delta = \sigma_\gamma^2/(\sigma_\gamma^2 + \sigma_\delta^2) = 1/(1 + r)$, where $r = \sigma_\delta^2/\sigma_\gamma^2$. This implies that how much we should increase our belief about δ as $\hat{\delta}$ increases does not depend on the absolute variance of the prior beliefs, but the relative variance. If we are relatively more uncertain about the treatment effect, m_δ is high, and when we are relatively uncertain about the bias, then m_δ is low.

Second-moment learning. The posterior variance of δ tells us whether there is second-moment learning about the treatment effect when observing $\hat{\delta}$:

$$\bar{\sigma}_\delta^2 = \sigma_\delta^2 \frac{(1 - \rho^2)\sigma_\gamma^2}{\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2}. \tag{5}$$

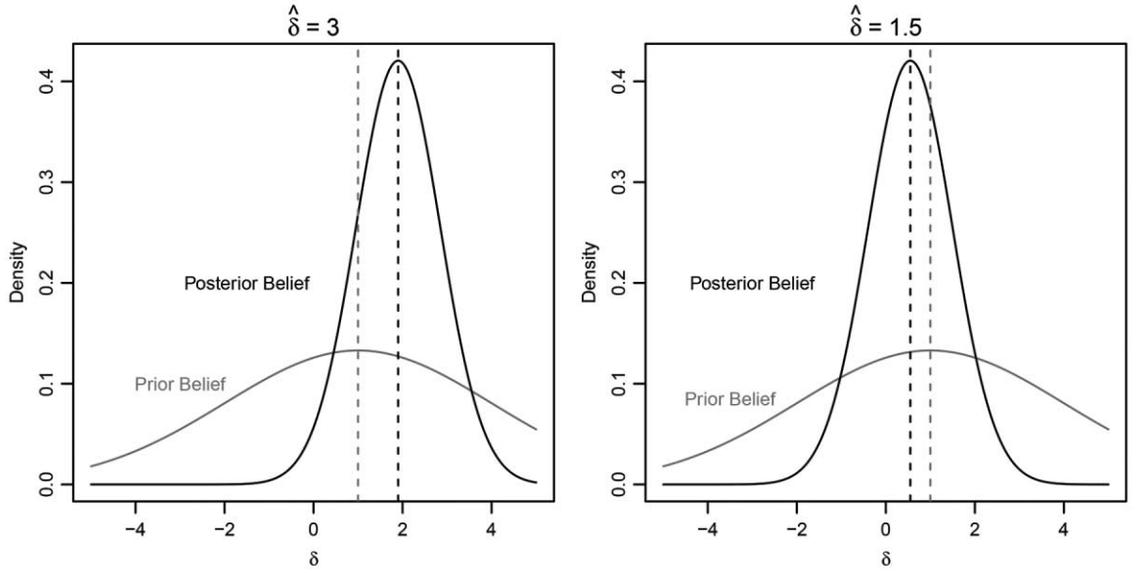

Figure 1. Two examples of updating. Simulation values are $\sigma_\delta^2 = 3$, $\sigma_\gamma^2 = 1$, $\mu_\delta = \mu_\gamma = 1$, $\rho = 0$. Biased estimate $\hat{\delta} = 3$ (*left*) and 1.5 (*right*). Color version available as an online enhancement.

It is immediate from the law of total variance that $\bar{\sigma}_\delta^2 \leq \sigma_\delta^2$. Further, unless $\rho = -\sigma_\delta/\sigma_\gamma$, this inequality is strict.[3] In words, other than a knife-edged case, if there are priors, then there is learning, 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. What this means in practice is that, perhaps surprisingly, so long as the researcher holds a proper prior belief on δ and γ she will have more precise beliefs about δ from seeing $\hat{\delta}$ from any research design than she would from not seeing $\hat{\delta}$. But if the researcher is unwilling to place any information in the prior belief about σ_γ or σ_γ, it follows that $\sigma_\gamma, \sigma_\delta \to \infty$ and as a result the posterior variances $\bar{\sigma}_\delta^2, \bar{\sigma}_\gamma^2 \to \infty$. In this case, the researcher's beliefs about δ are no more precise upon observing any $\hat{\delta}$.

What happens if the prior variances are not finite? For any given level of uncertainty in the prior belief about the treatment effect δ, with complete uncertainty about γ ($\sigma_\gamma \to \infty$, i.e., there is no prior information about the bias term), then $m_\delta \to 0$. The posterior distribution about δ in this case is normally distributed with mean 0 and variance $(1 - \rho^2)\sigma_\delta^2$. 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 et al. (2014), who argue that learning from observational research—where the bias term is unknown and represented as having infinite variance—is an "illusion." However, as argued above, the assumption of infinite variance on this prior is rarely, if ever, tenable.

Revisiting the motivating example

We return now to our motivating example in which we seek to learn if authoritarian elections reduce repression (meaning that δ is positive). We observe an estimate $\hat{\delta}$ that is the difference in the physical integrity rights of citizens of electoral authoritarian regimes and the physical integrity rights of citizens of nonelectoral authoritarian regimes. We would like to interpret this as an estimate of δ, the effect of elections on physical integrity rights. But we suspect that this represents both the effect of elections and some bias emerging from the fact that elections may be held by regimes that do not need to repress their citizens. How shall we proceed to learn about δ from $\hat{\delta}$?

Our approach is to place priors beliefs on δ and calculate our posterior estimate of μ_δ and $\bar{\sigma}_\delta^2$. Let us imagine a scenario in which we are relatively more uncertain about the effect of elections ($\sigma_\delta^2 = 3$) than we are about bias ($\sigma_\gamma^2 = 1$), 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 effect of elections upward—and increased the precision of that estimate—based on an estimated coefficient that is substantially larger than our prior belief and the bias term. Where $\hat{\delta} = 1.5$, by contrast, our posterior mean belief

3. The inequality $\sigma_\delta^2 \leq \bar{\sigma}_\delta^2$ holds if and only if $[(1 - \rho^2)\sigma_\gamma^2]/(\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2) \leq 1$. Rearranging, this is met if and only if $(\sigma_\delta - \rho\sigma_\gamma)^2 \geq 0$. This holds with equality when $\rho = -\sigma_\delta/\sigma_\gamma$ and strictly otherwise. As shown in app. A.2, this condition is met precisely when $\hat{\delta}$ and δ are independent.

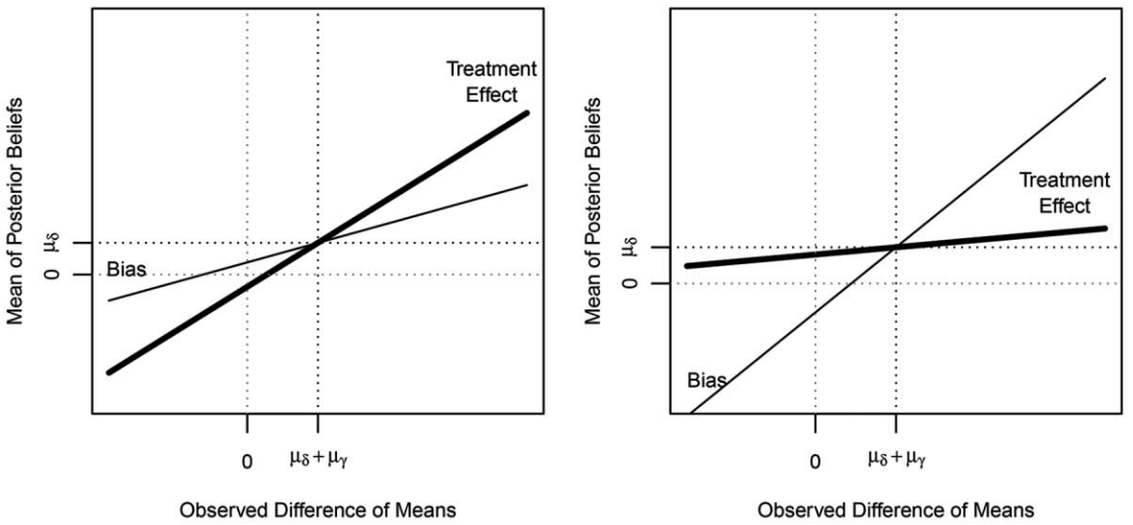

Figure 2. Updating beliefs about treatment and bias from the observed treatment effect. The thick line denotes mean posterior beliefs about the treatment effect, and the thin line denotes mean posterior beliefs about bias. In the left panel, $\sigma_\delta^2 = 3$, $\sigma_\gamma^2 = 1$, $\mu_\delta = \mu_\gamma = 1$, $\rho = 0$. In the right panel, the means and correlation are the same but $\sigma_\delta^2 = 1$ and $\sigma_\gamma^2 = 3$. Color version available as an online enhancement.

about the effect of elections 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 effect of elections since it is larger than the mean belief about bias.

In figure 2 we show more generally how posterior beliefs about the effects of election and the presence of 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. The black dotted lines correspond to the mean of the prior belief about the effect of elections (horizontal) and difference of means (vertical). When the observed difference of means between electoral and nonelectoral regimes is exactly equal to the prior $\mu_\gamma + \mu_\delta$ (here, 2), the mean of the posterior belief about the effect of elections is unchanged (as is the belief about bias). Though, importantly, there is still learning in this case, as the variance of the posterior belief goes down (i.e., there is second-moment learning).

For larger estimated effects of elections, 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 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 that there is more learning about this parameter. Still, higher differences of means lead to a higher belief about the treatment effect.

PRACTICAL CONSIDERATIONS

The model in the previous section is highly stylized, abstracting away from many practicalities of the research process. While this helps make our key points in a simple fashion, it could also raise concerns that these points are not applicable to actual empirical research. In this section, we apply our approach to more realistic settings, shedding light on several contentious questions in contemporary research methods: aggregating across multiple studies/estimates, the file drawer problem and other selective reporting, the "artificiality" of experiments, external validity and designs based on subsamples, and the possibility of serious research malpractice (purposeful or not).

What if there is sampling error or there are multiple estimates?

A first step toward addressing common issues in applied work is to ask how (1) sampling error and (2) multiple studies/estimates affect what we learn from potentially biased estimates.

The analysis with normal distributions can easily incorporate sampling error, so $\hat{\delta} = \delta + \gamma + \epsilon$, where ϵ is normally distributed with mean 0 and variance σ_ϵ^2.[4] By an analogous calculation (assuming that sampling error is uncorrelated with the treatment effect and bias in the prior) the posterior

4. For simplicity and ease of comparison to the baseline we deviate from standard Bayesian models of a difference of means or regression which would specify a prior belief of the variance of the error term and estimate that from the data. Even in this context the unit principle still holds, and we see no reason why a more complicated learning model would overturn our main claims.

beliefs about the treatment effect and bias can still be expressed as written in (3)–(4), but with

$$m_\delta = \frac{\sigma_\delta^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2 + \sigma_\epsilon^2}, \quad (6)$$

$$m_\gamma = \frac{\sigma_\gamma^2 + \rho\sigma_\delta\sigma_\gamma}{\sigma_\delta^2 + 2\rho\sigma_\delta\sigma_\gamma + \sigma_\gamma^2 + \sigma_\epsilon^2}. \quad (7)$$

The posterior belief about how much sampling error was in this estimate is $m_\epsilon(\hat{\delta} - \mu_\delta - \mu_\gamma)$ where $m_\delta + m_\gamma + m_\epsilon = 1$. 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.

What happens if we are able to reduce sampling error by doing larger studies or by conducting multiple studies with the same research design on different samples? The answer to the former question is straightforward: as $\sigma_\epsilon \to 0$, the right-hand sides of equations (6)–(7) become the same as the m terms in equations (3)–(4). In other words, we approach the case with no sampling error.

A similar principle holds if we observe multiple studies with the same research design, which we can capture in our modeling framework by assuming that they have the same bias term. Formally, assume we observe a series of estimates $\hat{\delta}_1, \ldots, \hat{\delta}_n$, given by

$$\hat{\delta}_i = \delta + \gamma + \epsilon_i,$$

where each ϵ_i is normally distributed with mean 0 and standard deviation $\sigma_{\epsilon,i}$ independent from all other random variables. The average of these estimates is then

$$\sum_{i=1} \hat{\delta}_i / n = \delta + \gamma + \frac{\sum_i \epsilon_i}{n}.$$

When n is large, the variance of $\left(\sum_i \epsilon_i\right)/n$ approaches 0. This returns us to the case outlined above, with no sampling error. An implication of this result is that once we have minimized sampling error, there are only incremental benefits to conducting more studies with the same research design even though we are still uncertain about the true treatment effect.

If our studies vary not just in the sampling error but also in the bias, however, then additional studies will help. Clearly, if we observe n estimates of the form $\delta + \gamma_i$ where the γ_i's are independent and normally distributed, we can perfectly learn γ as $n \to \infty$.[5] The implication of this result is that multiple biased studies that are biased in different ways can help us to learn about causal effects. This conclusion provides an argument to justify a folk belief among empiricists: that employing multiple identification strategies within a single study should give us more confidence that an estimated causal effect is real. Note, however, that this result is driven by the assumption that the γ_i's are independent. Still, we show in appendix A.3 that the general principle that we learn more from multiple estimates where the bias terms are less correlated holds under reasonable conditions.

5. The average of the demeaned γ_i's is normally distributed with mean 0 and a variance that approaches zero as $n \to \infty$.

What can we learn from reported research?

We interpret the model that we have presented above as explaining what an individual researcher should learn upon producing an estimate about a causal quantity of interest, given her or his prior beliefs about that causal quantity and the research design used to produce that estimate. But it highlights the general problem of how to learn from research findings of any sort when the consumer of the findings has beliefs about the bias inherent in the research design. A wealth of recent research on publication bias (see, e.g., Franco, Malhotra, and Simonovits [2014] and citations therein) documents that statistically insignificant results are less likely to be published than statistically significant ones. Can our approach help us to understand what an observer should learn from others' published research? Using our model to analyze what one should learn from reading published (or unpublished) estimates by other researchers requires a more sophisticated model of the process by which results are reported, which we outline below.

If researchers reported the results of every estimate they ran in the process of producing a paper, this would not create any conceptual difficulties in translating our results so far, as the reader of their study (or observer of their presentation) would have the exact same information the researcher does. Of course, as discussed above, researchers are selective in what they report, for both good reasons (e.g., to not overwhelm the audience and as the researcher plausibly has more information about which specifications are most reliable) and for less defensible reasons (e.g., selectively choosing the "strongest" results).

The challenge posed to our framework is that the audience for research only observes a subset of estimates of δ, and which estimates get released may depend on choices made by the researcher. To explore how this "filtering" process affects what we should learn from reported results, we analyze two processes of this form.

File drawer. First, suppose a researcher runs one estimate as in the main model, but only reports it to the audience if $\hat{\delta} \geq \bar{\delta}$ for some $\bar{\delta} \in \mathbb{R}$. If $\hat{\delta} < \bar{\delta}$, the researcher reports nothing. A natural motivation for this is that only sufficiently positive

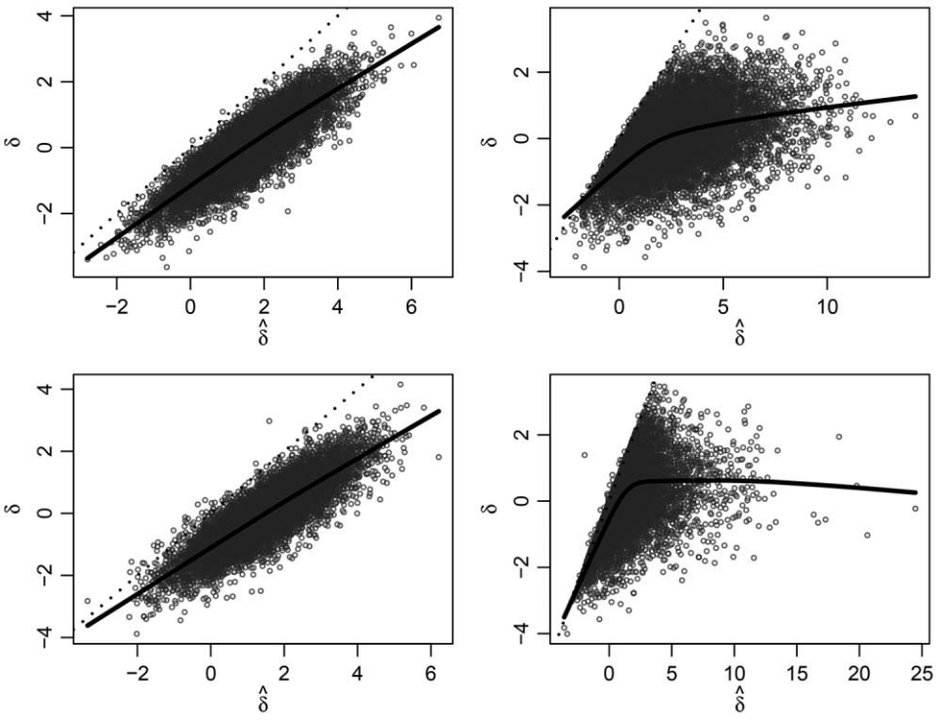

Figure 3. Simulations of selective reporting. In all four panels, $\delta \sim \mathcal{N}(0,1)$. Each data point represents a simulation where $\hat{\delta} = \delta + \max\{\gamma_1, \ldots, \gamma_m\}$. In the top panels $m = 10$, and in the bottom panels m is one plus a Poisson random variable with mean 9. In the left panels $\gamma_i \sim \mathcal{N}(0,1)$, and in the top right panel $\gamma_i \sim \mathcal{N}(0, \sigma_i)$ where σ_i is drawn from an exponential distribution with rate parameter 1.

result support the preferred theory of the researcher, or that only sufficiently positive results are "interesting" enough to be published.[6] In this case, the "signal" observed by the audience is

$$s = \begin{cases} \hat{\delta} & \text{if } \hat{\delta} \geq \bar{\delta} \\ \varnothing & \text{if } \hat{\delta} < \bar{\delta} \end{cases}, \tag{8}$$

where $s = \varnothing$ means "not writing the paper" (or perhaps the paper is not published).

We can now ask what the audience should learn about the ATET from s. Perhaps surprisingly, given this simplified process, the answer is the same as above, conditional on observing $s = \hat{\delta}$. Intuitively, once the result is reported, the audience has the same information as the researcher.

Inferences differ when the audience observes $s = \varnothing$. From this, the audience should infer that $\hat{\delta} < \bar{\delta}$, and as a result will tend to think that $\hat{\delta}$, and hence δ, is low. Of course, this inference requires the audience to know that the estimate has been produced, just not reported.

Specification search. In reality, researchers rarely estimate only one model and may choose what to report from among the results of many analyses. In this case, learning will prove elusive without further information about the specification search. To see why, we introduced a stylized example that allows us to provide structure on the learning problem. Suppose the researcher observes m estimates of the form:

$$\hat{\delta}_i = \delta + \gamma_i, \tag{9}$$

and reports $\hat{\delta}^{\max} = \max\{\hat{\delta}_1, \ldots, \hat{\delta}_m\}$. As with the file drawer example described previously, this could correspond to a case where more positive estimates are more consistent with the author's theory, or because editors and referees only agree to publish strong positive results. Of course, most applied research publishes more than one specification, but we simplify to make the argument.

The above considerations imply that the researcher will report the estimate with the most positive bias $\gamma^{\max} = \max\{\gamma_1, \ldots, \gamma_m\}$. If the audience happened to know how many estimates the researcher observed (m), and all of the γ_i's are independent and identically distributed (i.i.d.) random variables, some analytic progress can be made in characterizing the distribution of $\hat{\delta}^{\gamma}$ and $\gamma^{\max}$ using extreme value distributions. However, if the audience is uncertain about m, or if different specifications have different bias distributions, useful analytic characterizations prove elusive.

To illustrate what can be learned, we present in figure 3 some simulations of how various kinds of specification searches and reporting problems affect the relationship between the

6. Of course, this case differs from the standard problem of selecting results based on p-values, but for fixed standard errors, there is a one-to-one mapping from p-values to coefficient estimates.

estimates $\hat{\delta}$ that the audience observes and true causal effects δ about which they wish to learn. In this set of simulations, true causal effects δ are drawn from a normal distribution with $\mu_\delta = 0$ and $\sigma_\delta = 1$. Figure 3 presents four simulations. In each case, the x-axis is the reported $\hat{\delta}^{\max}$, and the y-axis is the true treatment effect δ. To read each figure, look first to the x-axis and ask, At this reported estimate, what does the distribution of true treatment effects look like?

In the top panels of figure 3, the number of estimates observed by the researcher is 10, while in the bottom panels it is a random variable with mean 10 (see the figure caption for more details). In other words, the top panels correspond to a case where the audience knows how many estimates were produced by the researcher, and in the bottom panels the audience does not know how many estimates were produced by the researcher. In the left panels, each bias term γ_i is normally distributed with mean 0 and variance 1, and in the right panels the bias terms have heterogeneous variances. That is, in the right-hand panels, some estimates contain more potential for bias than others. The dotted line corresponds to a 45 degree line (where the data would lie if the estimate $\hat{\delta}$ were equal to the true parameter δ) and the solid curve is a lowess fit.

We begin with the top left panel, where the "filter process" is "known" and bias terms are drawn from a normal distribution. Here, estimates tend to be larger than the treatment effect, but there is still a clear linear relationship between the two. So, compared to our baseline case, the audience should do some "subtraction" of the observed estimate to come up with a best guess of the treatment effect. But the principle that larger estimates lead to larger posterior beliefs about the treatment effect still holds. When we move to the bottom left panel, we discover that uncertainty about the number of specifications does not dramatically change what can be learned; the true estimates are just slightly more dispersed for any reported estimate $\hat{\delta}$.[7]

The right-hand panels show us, however, that uncertainty about the variance in the bias terms makes learning much harder. In the top right panel, the best fit (analogous to the mean of the posterior distribution) becomes nonlinear in $\hat{\delta}$. This is likely because larger estimates $\hat{\delta}$ tend to be driven by an estimate with an unusually extreme bias term. Another consequence of this fact is that as $\hat{\delta}$ gets large, there is more variance in the underlying real treatment effect—so we actually observe less second-moment learning. It nevertheless still follows that seeing higher treatment effects should lead the observer to think that the true causal effect is larger but also to be more suspicious that the reported result is driven by a bad research design. The bottom right panel presents the worst-case scenario, with uncertainty both about the distribution of bias and the number of estimates observed by the researcher. Here, the lowess curve eventually bends downward, meaning that the audience no longer can infer that larger reported estimates imply larger true causal effects.

The summary message from this exercise is that the nonrandom (and likely strategic) reporting of results leaves audiences with yet another layer of uncertainty when making inferences about causal effects. Like in a standard analysis, audiences are unsure about whether the estimate is subject to bias and how much it is contaminated by sampling error. The additional layer is that the audience is uncertain about what process led the estimate from the researcher's computer to the paper or presentation where it is observed.

Preregistration is commonly viewed as one way to address publication bias, helping audiences to learn better from published results. Our analysis helps to clarify how preregistration helps in this regard: it eliminates uncertainty about the number of analyses and where each result falls in the distribution $\{\hat{\delta}_1, \ldots, \hat{\delta}_m\}$, but not their bias in the analyses. Preregistration—at least when combined with a commitment to publish all analyses—thereby makes it possible to avoid the inferential problems found in the bottom right quadrant of figure 3. With certainty about that process, audiences can learn about δ using the approaches we have outlined above.

When should we learn more from designs 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

7. This is primarily because modest changes to the number of estimates do not lead to large changes in the shape of $\gamma^{\max}$. The maximum of i.i.d. normal random variables is roughly normally distributed, and the mean of this maximum increases in the number of variables but with quickly diminishing returns.

to take the treatment by the instrument). Many other kinds of natural experiments have a similar flavor.

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. Here we show how to take a belief about this similarity (however derived) and produce a posterior belief about the true parameter of interest.

Suppose we have a subsample estimate:

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

If γ^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 δ^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 δ then it is only valuable if we have a strong sense of the relationship between δ^s and δ.

To know what the researcher learns about δ 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 μ_{δ^s} and μ_{γ^s} be the priors means of the subsample properties, with variances $\sigma^2_{\delta^s}$ and $\sigma^2_{\gamma^s}$. The update on δ conditional on $\hat{\delta}^s$ is normal with mean:

$$\bar{\mu}_\delta = \mu_\delta + m_{\delta^s}(\hat{\delta}^s - \mu^s_\delta - \mu^s_\gamma), \tag{10}$$

where

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

and the posterior variance of the estimate of δ is

$$\sigma^2_\delta - \frac{\mathrm{Cov}(\hat{\delta}^s, \delta)^2}{\mathrm{Var}(\hat{\delta})} = \sigma^2_\delta - \frac{(\rho_{\delta,\delta^s}\sigma_\delta\sigma_{\delta^s} + \rho_{\delta,\gamma^s}\sigma_\delta\sigma_{\gamma^s})^2}{\sigma^2_{\delta^s} + 2\rho_{\delta^s,\gamma^s}\sigma_{\delta^s}\sigma_{\gamma^s} + \sigma^2_{\gamma^s}}, \tag{11}$$

where $\rho_{x,y}$ is the prior correlation between x and y. (See app. A.4 for the full derivation.) The subsample estimate is preferred to the full sample $\hat{\delta}$ in the sense of doing a better job of shrinking the posterior variance of δ if and only if

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

An instructive special case is if there is no bias in the subsample, that is, $\sigma_{\gamma^s} = 0$. If so, then $\bar{\sigma}^2_\delta$ simplifies to $(1 - \rho^2_{\delta,\delta^s})\sigma^2_\delta$. If, further, $\rho_{\delta,\gamma} = 0$, then the subsample is more informative about the treatment effect than the full sample if and only if

$$\rho^2_{\delta,\delta^s} \geq \frac{\sigma^2_\delta}{\sigma^2_\delta + \sigma^2_\gamma}. \tag{12}$$

As long as the prior variances on γ and δ are finite, the right-hand side of equation (12) 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 ($\rho^2_{\delta,\delta^s} = 0$), then the full sample is always more informative. 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, that is, when the full sample difference of means is primarily informative about the treatment effect.

When do we learn more from experiments with limited external validity?

One common source of criticism of experimental social science is that experimental conditions are inevitably artificial, producing estimates of quantities that have no practical relationship to the causal quantity of interest. For example, lab experiments with subjects who are primarily university students given explicit rules for how to make decisions are used to study deliberative democracy (Karpowitz, Mendelberg, and Shaker 2012), and 300 individuals from a neighborhood in Kampala play dictator games to study how ethnicity affects public goods provision (Habyarimana et al. 2007). In each of these studies, experimental conditions are precisely controlled in order to facilitate the estimation of treatment effects, but doing so inevitably means that the experiment is a highly artificial representation of the social realities it is meant to capture. The result is a trade-off between studying the actual phenomenon of interest without the benefit of experimental control and experimental designs that are unbiased but possibly too artificial to be useful.

Our approach to learning from observational research readily extends to these sorts of trade-offs between artificial but unbiased versus realistic but biased research designs. The logic is precisely the same as the logic we used to analyze the generalizability of subsamples. In equation (12), replace ρ^2_{δ,δ^s} with ρ^2_{δ,δ^e}, where δ^e is the treatment effect in an artificial but

unbiased experiment. Now, we can express the trade-off between an artificial experiment and a biased but realistic research design in terms of, first, the correlation between experimental and real world effects, and second, whether the variance in the treatment effect is more driven by uncertainty about the real world treatment effect than it is by bias. And if the experiment is simply unrelated to the real world ($\rho^2_{\delta,\delta^*} = 0$), we would never prefer it to a biased but realistic research design.

What if the study might be completely flawed? (And other alternative prior distributions)

Consider the following three relatively common scenarios. In the first, you observe a study with an identification strategy that hinges on a key assumption that is either true or not.[8] In the second, you observe a study where you have a strong, theoretically informed prior belief that the true causal effect is almost certainly zero.[9] In the third, you observe a presentation or read a paper where the reported treatment effect is so implausibly large in magnitude that you suspect the researcher made a large mistake that renders their estimate unrelated to the true effect, or even engaged in fraudulent practice.

All three of these examples are hard to fit into our main example with normal prior beliefs on the treatment effect and bias. However, they can all be tractably analyzed by assuming that the prior on one or both δ and γ are drawn from a normal mixture. We can capture an identifying assumption being true or not by writing the prior on γ as a normal mixture, equal to $\gamma = 0$ with some probability (if the identifying assumption holds) and drawn from a normal distribution with complementary probability (if the identifying assumption does not hold). Conversely, the example with a strong theoretical prior that the treatment effect may be zero corresponds to a normal mixture where $\delta = 0$ with some probability and is drawn from a normal distribution otherwise. The example with a mistake or fraud combines elements of both: either the study is honest and competent and δ and γ are drawn as in the main example, or there is a mistake and/or fraud and $\hat{\delta}$ is drawn from an unrelated distribution.

In appendix A.5, we analyze a model of learning where prior beliefs are drawn from normal mixtures such as these. The main result is that the size of the estimate can change beliefs about which element of the mixture it came from. One important implication of having a normal mixture on the bias term (cases 1 and 3) is that it is possible to break the monotonicity where larger estimates of $\hat{\delta}$ always lead to higher posterior beliefs about the treatment effect δ. Intuitively, larger observed estimates can increase the observer's skepticism that the study was unbiased/properly executed/not fraudulent.

In the case of a normal mixture prior on the treatment effect, however, there is still monotone learning where a higher difference of means leads to a higher belief about the treatment effect. The main difference from the standard normal model is that this learning can be highly nonlinear. For example, if the observer starts with a prior that the treatment effect is very likely zero, there will be minimal learning for typical returned differences of means. However, upon observing a very large difference of means, the belief that the true treatment effect is zero starts to become implausible, leading to large differences between prior and posterior beliefs about δ given $\hat{\delta}$.

Of course, normal mixtures are not the only alternative prior distributions to consider. For example, a more natural way to think about having extreme uncertainty may be to assume that δ and γ are drawn from (independent) Cauchy distributions with location parameters m_δ and m_γ. The analysis of this case, which appears in appendix A.5, is remarkably similar to our main case of a multivariate normal distribution.

DISCUSSION AND CONCLUSION

In this paper we have formulated the problem of learning about causal 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 use a Bayesian framework to ask what it means to learn about a causal effect and how prior beliefs about it and the bias allow us to update our posterior beliefs about that effect. Ignoring concerns about bias ("head in the sand") and claiming to learn nothing from studies potentially subject to bias ("staring at the sun") both represent extreme cases of learning that, we argue, do not reflect the actual beliefs that researchers have. We show that under more reasonable assumptions, we can learn something from biased research designs, though what we learn depends on prior beliefs about how bad the bias is.

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

8. For example, Kocher and Monteiro (2016) argue that Ferwerda and Miller's (2014) 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.

9. A prominent example here is Bem (2011) on extrasensory perception, which played a central role in uncovering the problems of p-hacking in psychology.

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 the observed 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 the section "Yes, You Have Priors" why we think that researchers almost always have priors about treatment effects and bias terms, even if they are unwilling to specify them precisely or even if we suspect that they have motivated reasons to raise concerns about bias. 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."

As in most areas of Bayesian analysis, 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 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 $\hat{\delta} > \mu_\gamma$. Researchers should therefore clearly state a "best guess" of how bad the bias is, as this will allow them to know in which direction they should update their beliefs.

To conclude, we briefly revisit the social aspects of scientific learning. Although our baseline approach focuses on individual learning rather than the strategic aspects of researcher behavior such as "p-hacking" and the "file drawer problem," we showed in the section "What Can We Learn from Reported Research" how we can use our approach in situations where an audience believes that researchers may have motivations to present particular results. We also suggested that researchers may have incentives to misrepresent their prior beliefs about bias, although we did not formally analyze how this affects the learning process. We view combining our approach with other recent papers on these topics as a promising direction for further research. The state of the art research in the econometrics of program evaluation approaches similar questions from different angles: Spiess (2018) provides guidelines that allow for specification searches where researchers may have preferences for obtaining particular outcomes, whereas Banerjee et al. (2017) explain why randomization is optimal when trying to confront a skeptical audience. Andrews and Kasy (2019) present several methods for estimating the probability of observing particular values of $\hat{\delta}$ given incentives to publish statistically significant results, which would in turn prove useful for updating beliefs about δ. And Fowler (2019) provides a method to adjust observed estimates for publication bias that could also incorporate prior beliefs about bias and measurement error. Further extensions in this direction will provide a more realistic normative account of the process of social learning from biased research designs, but our analysis on individual researcher learning is the first step in developing such a normative account.

ACKNOWLEDGMENTS

Thanks to Neal Beck, Matt Blackwell, Alex Coppock, Daniel de Kadt, Sean Gailmard, Andy Hall, Susan Hyde, Sabrina Karim, Mike Miller, Kevin Munger, Cyrus Samii, Keith Schnakenberg, Jas Sekhon, and three anonymous reviewers for comments and discussion.

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): 151–84.

Andrews, Isaac, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

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.

Banerjee, Abhijit, Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2017. "A Theory of Experimenters." Technical report, National Bureau of Economic Research, Cambridge, MA.

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.

Bueno De Mesquita, Ethan. 2010. "Regime Change and Revolutionary Entrepreneurs." *American Political Science Review* 104 (3): 446–66.

Calvert, Randall L. 1985. "The Value of Biased Information: A Rational Choice Model of Political Advice." *Journal of Politics* 47 (2): 530–55.

DellaVigna, Stefano, and Devin Pope. 2016. "Predicting Experimental Results: Who Knows What?" Technical report, National Bureau of Economic Research, Cambridge, MA.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 (2): 282–93.

Fearon, James D. 1999. "Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance." In Adam Przeworski, Susan C. Stokes, and Bernard Manin, eds., *Democracy, Accountability, and Representation*. Cambridge: Cambridge University Press.

Ferwerda, Jeremy, and Nicholas L. Miller. 2014. "Political Devolution and Resistance to Foreign Rule: A Natural Experiment." *American Political Science Review* 108 (3): 642–60.

Fowler, Anthony. 2019. "Correcting Point Estimates for Publication Bias." Unpublished manuscript.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–5.

Gandhi, Jennifer. 2008. *Political Institutions under Dictatorship*. New York: Cambridge University Press.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2014. "The Illusion of Learning from Observational Research." In Dawn Langan Teele, ed., *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven, CT: Yale University Press, 9–32.

Gill, Jeff, and Lee D. Walker. 2005. "Elicited Priors for Bayesian Model Specifications in Political Science Research." *Journal of Politics* 67 (3): 841–72.

Gustafson, Paul. 2015. *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*. Boca Raton, FL: Chapman & Hall/CRC.

Habyarimana, James, Macartan Humphreys, Daniel Posner, and Jeremy M. Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101 (4): 709–25.

Karpowitz, Christopher F., Tali Mendelberg, and Lee Shaker. 2012. "Gender Inequality in Deliberative Participation." *American Political Science Review* 106 (3): 533–47.

Kocher, Matthew A., and Nuno P. Monteiro. 2016. "Lines of Demarcation: Causation, Design-Based Inference, and Historical Research." *Perspectives on Politics* 14 (4): 952–75.

Little, Andrew T. 2017. "Are Non-competitive Elections Good for Citizens?" *Journal of Theoretical Politics* 29 (2): 214–42.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80 (2): 319–23.

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

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6): 1311–34.

McCandless, Lawrence C., Paul Gustafson, and Adrian Levy. 2007. "Bayesian Sensitivity Analysis for Unmeasured Confounding in Observational Studies." *Statistics in Medicine* 26 (11): 2331–47.

Oster, Emily. 2017. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics* 37 (2): 187–204.

Pepinsky, Thomas B. 2014. "The Institutional Turn in Comparative Authoritarianism." *British Journal of Political Science* 44 (3): 631–53.

Rivera, Mauricio. 2016. "Authoritarian Institutions and State Repression: The Divergent Effects of Legislatures and Opposition Parties on Personal Integrity Rights." *Journal of Conflict Resolution* 61 (10): 2183–2207.

Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. New York: Springer.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *Journal of Politics* 78 (3): 941–55.

Spiess, Jann. 2018. "Optimal Estimation When Researcher and Social Preferences Are Misaligned." Working paper, Harvard University, Department of Economics.

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