3 RCTs versus Observational Research

Assessing the Trade-Offs

Christopher H. Achen

3.1 Introduction

Experiments of all kinds have once again become popular in the social sciences (Druckman et al. 2011). Of course, psychology has long used them. But in my own field of political science, and in adjacent areas such as economics, far more experiments are conducted now than in the twentieth century (Jamison 2019). Lab experiments, survey experiments, field experiments – all have become popular (for example, Karpowitz and Mendelberg 2014; Mutz 2011; and Gerber and Green 2012, respectively; Achen 2018 gives an historical overview).

In political science, much attention, both academic and popular, has been focused on field experiments, especially those studying how to get citizens to the polls on election days. Candidates and political parties care passionately about increasing the turnout of their voters, but it was not until the early twenty-first century that political campaigns became more focused on testing what works. In recent years, scholars have mounted many field experiments on turnout, often with support from the campaigns themselves. The experiments have been aimed particularly at learning the impact on registration or turnout of various kinds of notifications to voters

I thank participants at the authors' workshop (at which drafts of chapters now comprising this volume were presented) for helpful advice. Sue Stokes and Larry Bartels provided timely suggestions. The responsibility for remaining errors and misjudgments remains with me.

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

that an election was at hand. (Green, McGrath, and Aronow 2013 reviews the extensive literature.)

Researchers doing randomized experiments of all kinds have not been slow to tout the scientific rigor of their approach. They have produced formal statistical models showing that an RCT is typically vastly superior to an observational (nonrandomized) study. In statistical textbooks, of course, experimental randomization has long been treated as the gold standard for inference, and that view has become commonplace in the social sciences. More recently, however, critics have begun to question this received wisdom. Cartwright (2007a, 2017, Chapter 2 this volume) and her collaborators (Cartwright and Hardie 2012) have argued that RCTs have important limitations as an inferential tool. Along with Heckman and Smith (1995), Deaton (2010) and others, she has made it clear what experiments can and cannot hope to do.

So where did previous arguments for RCTs go wrong? In this short chapter, I take up a prominent formal argument for the superiority of experiments in political science (Gerber et al. 2014). Then, building on the work of Stokes (2014), I show that the argument for experiments depends critically on emphasizing the central challenge of observational work – accounting for unobserved confounders – while ignoring entirely the central challenge of experimentation – achieving external validity. Once that imbalance is corrected, the mathematics of the model leads to a conclusion much closer to the position of Cartwright and others in her camp.

3.2 The Gerber–Green–Kaplan Model

Gerber, Green, and Kaplan (2014) make a case for the generic superiority of experiments, particularly field experiments, over observational research. To support their argument, they construct a straightforward model of Bayesian inference in the simplest case: learning the mean of a normal (Gaussian) distribution. This mean might be interpreted as an average treatment effect across the population of interest if everyone were treated, with heterogeneous treatment effects distributed normally. Thus, denoting the treatment-effects random variable by X_t and the population variance of the treatment effects by σ_t^2, we have the first assumption:

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

$$X_t \sim N(\mu, \sigma_t^2) \tag{1}$$

Gerber et al. implicitly take σ_t^2 to be known; we follow them here.[2]

In Gerber et al. (2014)'s setup, there are two ways to learn about μ. The first is via an RCT, such as a field experiment. They take the view that estimation of population parameters by means of random sampling is analogous to the estimation of treatment effects by means of randomized experimentation (Gerber et al. 2014, 32 at fn. 8). That is, correctly conducted experiments are always unbiased estimates of the population parameter.

Following Gerber et al.'s mathematics but making the experimental details a bit more concrete, suppose that the experiment has a treatment and a control group, each of size n, with individual outcomes distributed normally and independently: $N(\mu, \sigma_e^2/2)$ in the experimental group and $N(0, \sigma_e^2/2)$ in the control group. That is, the mathematical expectation of outcomes in the treatment group is the treatment effect μ, while the expected effect in the control group is 0. We assume that the sampling variance is the same in each group and that this variance is known. Let the sample means of the experimental and control groups be $\overline{x}_e$ and $\overline{x}_c$ respectively, and let their difference be $\widehat{\mu}_e = \overline{x}_e - \overline{x}_c$.

Then, by the textbook logic of pure experiments plus familiar results in elementary statistics, the difference $\widehat{\mu}_e$ is distributed as:

$$\widehat{\mu}_e \sim N(\mu, \sigma_e^2/n) \tag{2}$$

which is unbiased for the treatment effect μ. Thus, we may define a first estimate of the treatment effect by $\widehat{\mu}_e = \overline{x}_e - \overline{x}_c$: It is the estimate of the treatment effect coming from the experiment. This is the same result as in Gerber et al. (2014, 12), except that we have spelled out here the dependence of the variance on the sample size.

Next, Gerber et al. assume that there is a second source of knowledge about μ, this time from an observational study with m independent observations, also independent of the experimental observations. Via regression or other statistical methods, this study generates a normally distributed estimate of the treatment effect μ, with known sampling variance σ_o^2/m. However, because the methodology is not experimental, Gerber et al. (2014, 12–13) assume that the effect is estimated with confounding, so that its expected value is distorted by a bias term β. Hence, the estimate from the observational study $\widehat{\mu}_o$ is distributed as:

$$\widehat{\mu}_o \sim N(\mu + \beta, \sigma_o^2/m) \tag{3}$$

We now have two estimates, $\widehat{\mu}_e$ and $\widehat{\mu}_o$, and we want to know how to combine them. One can proceed by constructing a minimum-mean-squared error estimate in a classical framework, or one can use Bayesian methods. Since both approaches give the same result in our setup and since the Bayesian logic is more familiar, we follow Gerber et al. in adopting it. In that case, we need prior distributions for each of the unknowns.

With all the variances assumed known, there are just two unknown parameters, μ and β. An informative prior on μ is not ordinarily adopted in empirical research. At the extreme, as Gerber et al. (2014, 15) note, a fully informative prior for μ would mean that we already knew the correct answer for certain and we would not care about either empirical study, and certainly not about comparing them. Since our interest is in precisely that comparison, we want the data to speak for themselves. Hence, we set the prior variance on μ to be wholly uninformative; in the usual Bayesian way we approximate its variance by infinity.[1]

The parameter β also needs a prior. Sometimes we know the likely size and direction of bias in an observational study, and in that case we would correct the observational estimate by subtracting the expected size of the bias, as Gerber et al. (2014, 14) do. For simplicity here, and because it makes no difference to the argument, we will assume that the direction of the bias is unknown and has prior mean zero, so that subtracting its mean has no effect. Then the prior distribution is:

$$\beta \sim N(0, \sigma_\beta^2) \tag{4}$$

Here σ_β^2 represents our uncertainty about the size of the observational bias. Larger values indicate more uncertainty. Standard Bayesian logic then shows that our posterior distribution for the observational study on its own is $\widehat{\mu}_{op} = N(\mu, \sigma_o^2/m + \sigma_\beta^2)$.

Now, under these assumptions, Bayes' Theorem tells us how to combine the observational and experimental evidence, as Gerber et al. (2014, 14) point out. In accordance with their argument, the resulting

[1] Without this assumption, the Bayesian treatment estimate would differ from the minimum mean squared error estimate.

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

combined or aggregated estimate $\widehat{\mu}_a$ is a weighted average of the two estimates $\widehat{\mu}_{op}$ and $\widehat{\mu}_e$:

$$\widehat{\mu}_a = p\widehat{\mu}_{op} + (1 - p)\widehat{\mu}_e \tag{5}$$

where p is the fraction of the weight given to the observational evidence, and

$$p = \frac{\sigma_e^2/n}{\sigma_e^2/n + \sigma_o^2/m + \sigma_\beta^2} \tag{6}$$

This result is the same as Gerber et al.'s, except that here we had no prior information about μ, which simplifies the interpretation without altering the implication that they wish to emphasize.

That implication is this: Since σ_e^2, σ_o^2, n, and m are just features of the observed data, the key aspect of p is our uncertainty about the bias term β, which is captured by the prior variance σ_β^2. Importantly, Gerber et al. (2014, 15) argue that we often know relatively little about the size of likely biases in observational research. In the limit, they say, we become quite uncertain, and $\sigma_\beta^2 \to \infty$. In that case, obviously, $p \to 0$ in Equation (6), and the observational evidence gets no weight at all in Equation (5), not even if its sample size is very large.

This limiting result is Gerber et al.'s (2014, 15) Illusion of Observational Learning Theorem. It formalizes the spirit of much recent commentary in the social sciences, in which observational studies are thought to be subject to biases of unknown, possibly very large size, whereas experiments follow textbook strictures and therefore reach unbiased estimates. Moreover, in an experiment, as the sample size goes to infinity, the correct average treatment effect is essentially learned with certainty.[2] Thus, only experiments tell us the truth. The mathematics here is unimpeachable, and the conclusion and its implications seem to be very powerful. Gerber et al. (2014, 19–21) go on to demonstrate that under conditions like these, little or no resources should be allocated to observational research. We cannot learn anything from it. The money should go to field experiments such as those they have conducted, or to other experiments.

[2] That is, the posterior distribution collapses around the true treatment effect μ, or in classical terms, $\text{plim}\hat{\mu}_e = \mu$.

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

3.3 A Learning Theorem with No Thumb on the Scale

Gerber et al.'s Illusion of Observational Learning Theorem follows rigorously from their assumptions. The difficulty is that those assumptions combine jaundiced cynicism about observational studies with gullible innocence about experiments. As they make clear in the text, the authors themselves are neither unrelievedly cynical nor wholly innocent about either kind of research. But the logic of their mathematical conclusion turns out to depend entirely on their becoming sneering Mr. Hydes as they deal with observational research, and then transforming to kindly, indulgent Dr. Jekylls when they move to RCTs.

To see this, consider the standard challenge of experimental research: external validity, discussed in virtually every undergraduate methodology text (for example, Kellstedt and Whitten 2009, 75–76). Gerber et al. (2014, 22–23) mention this problem briefly, but they see it as a problem primarily for laboratory experiments because the inferential leap to the population is larger than for field experiments. The challenges that they identify for field experiments consist primarily in administering them properly. Even then, they suggest that statistical adjustments can often correct the biases induced (Gerber et al. 2014, 23–24). The flavor of their remarks may be seen in the following sentence:

> The external validity of an experiment hinges on four factors: whether the subjects in the study are as strongly influenced by the treatment as the population to which a generalization is made, whether the treatment in the experiment corresponds to the treatment in the population of interest, whether the response measure used in the experiment corresponds to the variable of interest in the population, and how the effect estimates were derived statistically. (Gerber et al. 2014, 21)

What is missing from this list are the two critical factors emphasized in the work of recent critics of RCTs: heterogeneity of treatment effects and the importance of context. A study of inducing voter turnout in a Michigan Republican primary cannot be generalized to what would happen to Democrats in a general election in Louisiana, where the treatment effects are likely to be very different. There are no Louisianans in the Michigan sample, no Democrats, and no general election voters. Hence, no within-sample statistical adjustments are available to accomplish the inferential leap. Biases of unknown magnitude remain, and these are multiplied when one aims to generalize to a national population as a whole. As Cartwright (2007a;

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

Chapter 2 this volume), Cartwright and Hardie 2012, Deaton (2010), and Stokes (2014) have spelled out, disastrous inferential blunders occur commonly when a practitioner of field experiments imagines that they work the way Gerber et al. (2014) assume that they work in their Bayesian model assumptions. Gerber et al. (2014, 32 at fn. 6) concede in a footnote: "Whether bias creeps into an extrapolation to some other population depends on whether the effects vary across individuals in different contexts." But that crucial insight plays no role in their mathematical model.

What happens in the Gerber et al. model when we take a more evenhanded approach? If we assume, for example, that experiments have a possible bias γ stemming from failures of external validity, then in parallel to the assumption about bias in observational research, we might specify our prior beliefs about external invalidity bias as normally and independently distributed:

$$\gamma = N(0, \sigma_\gamma^2) \tag{7}$$

Then the posterior distribution of the treatment estimate from the experimental research would be $\widehat{\mu}_{ep} = N(\mu, \sigma_e^2/n + \sigma_\gamma^2)$, and the estimate combining both observational and experimental evidence would become:

$$\widehat{\mu}_{ab} = q\widehat{\mu}_{op} + (1 - q)\widehat{\mu}_{ep} \tag{8}$$

where q is the new fraction of the weight given to the observational evidence, and

$$q = \frac{\sigma_e^2/n + \sigma_\gamma^2}{\sigma_e^2/n + \sigma_\gamma^2 + \sigma_o^2/m + \sigma_\beta^2} \tag{9}$$

A close look at this expression (or taking partial derivatives) shows that the weight given to observational and experimental evidence is an intuitively plausible mix of considerations.

For example, an increase in m (the sample size of the observational study) reduces the denominator and thus raises q; this means that, all else equal, we should have more faith in observational studies with more observations. Conversely, increases in n, the sample size of an experiment, raise the weight we put on the experiment. In addition, the harder that authors have worked to eliminate confounders in observational research (small σ_β^2), the more we believe them. And the fewer the issues with external validity in an experiment (small σ_γ^2), the more weight we put on the experiment. That is what follows from Gerber et al.'s line of analysis when all the

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

potential biases are put on the table, not just half of them. But, of course, all these implications have been familiar for at least half a century. Carried out evenhandedly, the Bayesian mathematics does no real work and brings us no real news.

Gerber et al. arrived at their Illusion of Observational Learning Theorem only by assuming away the problems of external validity in experiments. No surprise that experiments look wonderful in that case. But one could put a thumb on the other side of the scale: Suppose we assume that observational studies, when carefully conducted, have no biases due to omitted confounders, while experiments continue to have arbitrarily large problems with external validity. In that case, $\sigma_\beta^2 = 0$ and $\sigma_\gamma^2 \to \infty$. A look at Equations (8) and (9) then establishes that in that case, we get an Illusion of Experimental Learning Theorem: Experiments can teach us nothing, and no one should waste time and money on them. But of course, this inference is just as misleading as Gerber et al.'s original theorem.

Gerber et al. (2014, 11–12, 15, 26–30) concede that observational research sometimes works very well. When observational biases are known to be small, they see a role for that kind of research. But they never discuss a similar condition for valid experimental studies. Even in their verbal discussions, which are more balanced than their mathematics, they continue to write as if experiments had no biases: "experiments produce unbiased estimates regardless of whether the confounders are known or unknown" (Gerber et al. 2014, 25). But that sentence is true only if external validity is never a problem. Their theorem about the unique value of experimental work depends critically on that assumption. Alas, the last decade or two have taught us forcefully, if we did not know it before, that their assumption is very far from being true. Just as instrumental variable estimators looked theoretically attractive when they were developed in the 1950s and 1960s but often failed in practice (Bartels 1991), so too the practical limitations of RCTs have now come forcefully into view.

Experiments have an important role in political science and in the social sciences generally. So do observational studies. But the judgment as to which of them is more valuable in a particular research problem depends on a complex mixture of prior experience, theoretical judgment, and the details of particular research designs. That is the conclusion that follows from an evenhanded set of assumptions applied to the model Gerber et al. (2014) set out.

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

3.4 Conclusion

Causal inference of any kind is just plain hard. If the evidence is observational, patient consideration of plausible counterarguments, followed by the assembling of relevant evidence, can be, and often is, a painstaking process.[3] Faced with those challenges, researchers in the current intellectual climate may be tempted to substitute something that looks quicker and easier – an experiment.

The central argument for experiments (RCTs) is that the randomization produces identification of the key parameter. That is a powerful and seductive idea, and it works very well in textbooks. Alas, this modus operandi does not work nearly so well in practice. Without an empirical or theoretical understanding of how to get from experimental results to the relevant population of interest, stand-alone RCTs teach us just as little as casual observational studies. In either case, there is no royal road to secure inferences, as Nancy Cartwright has emphasized. Hard work and provisional findings are all we can expect. As Cartwright (2007b) has pungently remarked, experiments are not the gold standard, because there is no gold standard.

References

Achen, C. H. (2018) *Cycles in academic fashions: The case of experiments*. Mimeo, Department of Politics, Princeton University.

Achen, C. H. and Bartels, L. M. (2018) "Statistics as if politics mattered," *Journal of Politics*, 80 (4), 1438–1453.

Bartels, L. M. (1991) "Instrumental and quasi-instrumental variables," *American Journal of Political Science*, 35(3), 777–800.

Bross, I. D. J. (1960). "Statistical criticism," *Cancer*, 13(2), 394–400.

Cartwright, N. (2007a) *Hunting causes and using them*. New York: Cambridge University Press.

Cartwright, N. (2007b) "Are RCTs the gold standard?" *BioSocieties*, 2(1), 11–20.

Cartwright, N. (2017) "Single case causes: What is evidence and why" in Chao, H. and Reiss, J. (eds.) *Philosophy of science in practice: Nancy Cartwright and the nature of scientific reasoning*. New York: Springer International Publishing, pp. 11–24.

[3] Plausible counterarguments, but not just any speculative counterargument, a point not always understood by critics of observational evidence: see Bross (1960); Achen and Bartels (2018).

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press

Cartwright, N. and Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Deaton, A. (2010) “Instruments, randomization, and learning about development,” *Journal of Economic Literature*, 48(2), 424–455.

Druckman, J. N., Green, D. P., Kuklinski, J. H., and Lupia, A. (2011) *Cambridge handbook of experimental political science*. New York: Cambridge University Press.

Gerber, A. S. and Green, D. (2012). *Field experiments*. New York: Norton.

Gerber, A. S., Green, D. and Kaplan, E. H. (2014) “The illusion of learning from observational research” in Teele, D. L. (ed.) *Field experiments and their critics*. New Haven, CT: Yale University Press, pp. 9–32.

Green, D. P., McGrath, M. C., and Aronow, P. M. (2013) “Field experiments and the study of voter turnout,” *Journal of Elections, Public Opinion and Parties*, 23(1), 27–48.

Heckman, J. J. and Smith, J. A. (1995) “Assessing the case for social experiments,” *Journal of Economic Perspectives*, 9(2), 85–110.

Jamison, J. C. (2019) “The entry of randomized assignment into the social sciences,” *Journal of Causal Inference*, 7(1). http://dx.doi.org/10.1515/jci-2017-0025 (accessed January 21, 2020).

Karpowitz, C. F. and Mendelberg, T. (2014) *The silent sex: Gender, deliberation and institutions*. Princeton, NJ: Princeton University Press.

Kellstedt, P. M. and Whitten, G. D. (2009) *The fundamentals of political science research*. New York: Cambridge University Press.

Mutz, D. C. (2011) *Population-based survey experiments*. New York: Oxford University Press.

Stokes, S. C. (2014) “A defense of observational research,” in Teele, D. L. (ed.) *Field experiments and their critics*. New Haven, CT: Yale University Press, pp. 33–57.

Teele, D. L., ed. (2014) *Field experiments and their critics*. New Haven, CT: Yale University Press.

https://doi.org/10.1017/9781108688253.004 Published online by Cambridge University Press