This is a preprint of an article published in [Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. The Illusion of Learning from Observational Research. In Ian Shapiro, Rogers Smith, and Tarek Massoud, eds., *Problems and Methods in the Study of Politics*. New York: Cambridge University Press, pp. 251-73. Reprinted in *Field Experiments and their Critics*, Dawn Teele (ed.). New Haven: Yale University Press]. Pagination of this preprint may differ slightly from the published version.

12 The illusion of learning from observational research

Alan S. Gerber, Donald P. Green, and Edward H. Kaplan

Introduction

Empirical studies of cause and effect in social science may be divided into two broad categories, experimental and observational. In the former, individuals or groups are randomly assigned to treatment and control conditions. Most experimental research takes place in a laboratory environment and involves student participants, but several noteworthy studies have been conducted in real-world settings, such as schools (Howell and Peterson 2002), police precincts (Sherman and Rogan 1995), public housing projects (Katz, Kling, and Liebman 2001), and voting wards (Gerber and Green 2000). The experimental category also encompasses research that! examines the consequences of randomization performed by administrative agencies, such as the military draft (Angrist 1990), gambling lotteries (Imbens, Rubin, and Sacerdote 2001), random assignment of judges to cases (Berube 2002), and random audits of tax returns (Slemrod, Blumenthal, and Christian 2001). The aim of experimental research is to examine the effects of random variation in one or more independent variables.

Observational research, too, examines the effects of variation in a set of independent variables, but this variation is not generated through randomization procedures. In observational studies, the data generation process by which the independent variables arise is unknown to the researcher. To estimate the parameters that govern cause and effect, the analyst of observational data must make several strong assumptions about the statistical relationship between observed and unobserved causes of the dependent variable (Achen 1986; King, Keohane, and Verba 1994). To the extent that these assumptions are unwarranted, parameter estimates will be biased. Thus, observational research involves two types of

The authors are grateful to the Institution for Social and Policy Studies at Yale University for research support. Comments and questions may be directed to the authors at alan.gerber@yaie.edu, donald.green@yale.edu, or edward.kaplan@yale.edu.

uncertainty, the statistical uncertainty given a particular set of modeling assumptions and the theoretical uncertainty about which modeling assumptions are correct.

The principal difference between experimental and observational research is the use of randomization procedures. Obviously, random assignment alone does not guarantee that an experiment will produce unbiased estimates of causal parameters (cf. Cook and Campbell 1979, ch. 2, on threats to internal validity). Nor does observational analysis preclude unbiased causal inference. The point is that the risk of bias is typically much greater for observational research. This chapter characterizes experiments as unbiased and observational studies potentially biased, but the analytic results we derive generalize readily to situations in which both are potentially biased.

The vigorous debate between proponents of observational and experimental analysis (Cook and Payne 2002; Heckman and Smith 1995; Green and Gerber 2003; Weiss 2002) raises two meta-analytic questions. First, under what conditions and to what extent should we update our prior beliefs based on experimental and observational findings? Second, looking to the future, how should researchers working within a given substantive area allocate resources to each type of research, given the costs of each type of data collection?

Although these questions have been the subject of extensive discussion, they have not been addressed within a rigorous analytic framework. As a result, many core issues remain unresolved. For example, is the choice between experimental and observational research fundamentally static, or does the relative attractiveness of experimentation change depending on the amount of observational research that has accumulated up to that point in time? To what extent and in what ways is the tradeoff between experimental and observational research affected by developments in "theory" and in "methodology"?

The analytic results presented in this chapter reveal that the choice between experimental and observational research is fundamentally dynamic. The weight accorded to new evidence depends upon what methodological inquiry reveals about the biases associated with an estimation procedure as well as what theory tells us about the biases associated with our extrapolations from the particularities of any given study. We show that the more one knows *ex ante* about the biases of a given research approach, the more weight one accords the results that emerge from it. Indeed, the analytics presented below may be read as an attempt to characterize the role of theory and methodology within an observational empirical research program. When researchers lack prior information about the biases associated with observational research, they will

assign observational findings zero weight and will never allocate future resources to it. In this situation, learning is possible only through unbiased empirical methods, methodological investigation, or theoretical insight. These analytic results thus invite social scientists to launch a new line of empirical inquiry designed to assess the direction and magnitude of research biases that arise in statistical inference and extrapolation to other settings.

Assumptions and notation

Suppose you seek to estimate the causal parameter M. To do so, you launch two empirical studies, one experimental and the other observational. In advance of gathering the data, you hold prior beliefs about the possible values of *M.* Specifically, your prior beliefs about *M* are distributed normally with mean μ and variance σ^2 m. The dispersion of your prior beliefs (σ^2 *M*) is of special interest. The smaller σ^2 **м,** the more certain you are about the true parameter *M* in advance of seeing the data. An infinite σ^2 *M* implies that you approach the research with no sense whatsoever Of where the truth lies.

You now embark upon an experimental study. Before you examine the data, the central limit theorem leads you to believe that your estimator, X_e, will be normally distributed. Given that $M = m$ (the true effect turns out to equal *m)* and that random assignment of observations to treatment and control conditions renders your experiment unbiased, X_e is normal with mean ***m*** and variance $\sigma^2_{\boldsymbol{x_e}}$. As a result of the study, you will observe a draw from the distribution of X_e, the actual experimental value

In addition to conducting an experiment, you also gather observational data. Unlike randomized experimentation, observational research does not involve a procedure that ensures unbiased causal inference. Thus, before examining your observational results, you harbor prior beliefs about the bias associated with your observational analysis. Let *B* be the random variable that denotes this bias. Suppose that your prior beliefs about *B* are distributed normally with mean β and variance σ^2_B Again, smaller values of er^2_B indicate more precise prior knowledge about the nature of the observational study's bias. Infinite variance implies complete uncertainty.

Further, we assume that priors about *M* and *B* are independent. This assumption makes intuitive sense: there is usually no reason to suppose *ex ante* that one can predict the observational study's bias by knowing whether a causal parameter is large or small. It should be stressed, however, that independence will give way to a negative correlation once

the experimental and observational results become known.[1] The analytic results we present here are meant to describe what happens as one moves from prior beliefs to posterior views based on new information. The results can also be used to describe what happens after one examines an entire literature of experimental and observational studies. The precise sequence in which one examines the evidence does not affect our conclusions, but tracing this sequence does make the analytics more complicated. For purposes of exposition, therefore, we concentrate our attention on what happens as one moves from priors developed in advance of seeing the results to posterior views informed by all the evidence that one observes subsequently.

The observational study generates a statistical result, which we denote X_0 *(o* for observational). Given that $M = m$ (the true effect equals *m)* and $B = b$ (the true bias equals *b),* we assume that the sampling distribution of $\boldsymbol{X_0}$ is normal with mean $m + b$ and variance $\sigma^2 x_o$. In other words, the observational study produces an estimate (x_o) that may be biased in the event that b is not equal to *0.* Bias may arise from any number of sources, such as unobserved heterogeneity, errors in variables, and other well-known problems. The variance of the observational study ($a^2 x_0$) is a function of sample size, the predictive accuracy of the model, and other features of the statistical analysis used to generate the estimates.

Finally, we assume that given $M=m$ and $B = b$, the random variables X_e and $\boldsymbol{X_0}$ are independent. This assumption follows from the fact that the experimental and observational results do not influence each other in any way. In sum, our model of the research process assumes (1) normal and independently distributed priors about the true effect and the bias of observational research and (2) normal and independently distributed sampling distributions for the estimates generated by the experimental and observational studies. We now examine the implications of this analytic framework.

The joint posterior distribution of M and B

The first issue to be addressed is how our beliefs about the causal parameter M will change once we see the results of the experimental and observational studies. The more fruitful the research program, the more

1. This negative correlation results from the fact that the experiment provides an unbiased estimator of M, whereas the observational study provides an unbiased estimator of M + B. As we note below, once these findings become known, higher estimates of M from the experimental study imply lower values of B when the experimental result is subtracted from the observational result.

our posterior beliefs will differ from our prior beliefs. New data might give us a different posterior belief about the location of M, or it might confirm our prior belief and reduce the variance (uncertainty) of these beliefs.

Let fx(x) represent the normal probability density for random variable *X* evaluated at the point ***x,*** and let *fx\A* (x) be the conditional density for *X* evaluated at the point x given that the event *A* occurred. Given the assumptions above, the joint density associated with the compound event $M = m$, $X_e = x_e$, $B = b$, and $X_o = x_o$ is given by

$$f_{\mathrm{M}}(m) \times f_{\mathrm{x_e}|M=m}(x_e) \times f_{\mathrm{B}}(b) \times f_{\mathrm{x_o}|M=m,\, B=b}(x_o). \qquad (1)$$

What we want is the joint posterior distribution of M, the true effect, and B, the bias associated with the observational study, given the experimental and observational data. Applying Bayes' rule we obtain:

$$f_{\mathrm{M,\,B}|X_e=x_e, X_o=x_o}(m,b)$$

$$= \qquad (2)$$

Integrating over the normal probability distributions (cf. Box and Tiao 1973) produces the following result.

Theorem 1: The joint posterior distribution of *M* and *B* is bivariate normal with the following means, variances, and correlation.

The posterior distribution of M is normally distributed with mean given by

$$E(M \mid X_e = x_e, X_o = x_o) = p_1\mu + p_2 x_e + p_3(x_o - \beta)$$

and variance

$$\sigma^2_{M|x_e,x_o} = \frac{1}{\frac{1}{\sigma_M^2} + \frac{1}{\sigma_{x_e}^2} + \frac{1}{\sigma_B^2 + \sigma_{X_o}^2}},$$

where

$$p_1 = \frac{\sigma^2_{M|x_e,x_o}}{\sigma_M^2},\ p_2 = \frac{\sigma^2_{M|x_e,x_o}}{\sigma_{x_e}^2},\ \text{and}\ p_3 = \frac{\sigma^2_{M|x_e,x_o}}{\sigma_M^2+\sigma_{X_o}^2}$$

The posterior distribution of *B* is normally distributed with mean

$$E(B|X_e = x_e,\ X_o = x_o) = q\,1\beta + q_2(x_o - \mu) + q3(x_o - x_e)$$

and variance

$$\sigma^2_{B|x_e,x_o} = \frac{1}{\frac{1}{\sigma^2_B} + \frac{\frac{1}{\sigma^2_{X_o}}\left(\frac{1}{\sigma^2_M} + \frac{1}{\sigma^2_{X_e}}\right)}{\frac{1}{\sigma^2_M} + \frac{1}{\sigma^2_{X_e}} + \frac{1}{\sigma^2_{X_o}}}},$$

where

$$q_1 = \frac{\sigma^2_{B|x_e,x_o}}{\sigma^2_B},$$

$$q_2 = \frac{\sigma^2_{B|x_e,x_o}\left(\frac{1}{\sigma^2_M\sigma^2_{X_o}}\right)}{\frac{1}{\sigma^2_M} + \frac{1}{\sigma^2_{X_e}} + \frac{1}{\sigma^2_{X_o}}},$$

$$q_3 = \frac{\sigma^2_{B|x_e,x_o}\left(\frac{1}{\sigma^2_{X_e}\sigma^2_{X_o}}\right)}{\frac{1}{\sigma^2_M} + \frac{1}{\sigma^2_{X_e}} + \frac{1}{\sigma^2_{X_o}}},$$

The correlation between *M* and *B* after observing the experimental and observational findings is given by $\rho < 0$, such that

$$1 - p^2 = \frac{\frac{1}{\sigma^2_M} + \frac{1}{\sigma^2_{X_e}} + \frac{1}{\sigma^2_B + \sigma^2_{X_o}}}{\frac{1}{\sigma^2_M} + \frac{1}{\sigma^2_{X_e}} + \frac{1}{\sigma^2_{X_o}}}$$

This theorem reveals that the posterior mean is an average (since $P1 + P_2 + p_3 = 1$) of three terms: the prior expectation of the true mean effect (μ), the observed experimental value (x_e), and the observational value corrected by the prior expectation of the bias (x_O—β). This analytic result parallels the standard case in which normal priors are confronted with normally distributed evidence (Box and Tiao 1973). In this instance, the biased observational estimate is recentered to an unbiased estimate by subtracting off the prior expectation of the bias. It should be noted that such recentering is rarely, if ever, done in practice. Those who report observational results seldom disclose their'priors about the bias term,

let alone correct for it. In effect, researchers working with observational data routinely, if implicitly, assume that the bias equals zero and that the uncertainty associated with this bias is also zero.

To get a feel for what the posterior distribution implies substantively, it is useful to consider several limiting cases. If prior to examining the data one were certain that the true effect were μ, then $\sigma^2 M = 0$, $p1 = 1$, and $p_2 = p_3 — 0$. In this case, one would ignore the data from both studies and set $E(M \mid X_e = X_O = x_O) = \mu$. Conversely, if one had no prior sense of M or B before seeing the data, then $\sigma^2 M = \sigma^2 B = \infty$, $p1 = P3 = 0$, and $p2 = 1$, in which case the posterior expectation of M would be identical to the experimental result x_e. In the less extreme case one has some prior information about M such that $\sigma^2 M < \infty$, p_3 remains zero so long as one remains completely uninformed about the biases of the observational research. In other words, in the absence of prior knowledge about the bias of observational research, one accords it zero weight. Note that this result holds even when the sample size of the observational study is so large that a^2_{Xo} is reduced to zero.

For this reason, we refer to this, result as the Illusion of Observational Learning Theorem. If one is entirely uncertain about the biases of observational research, the accumulation of observational findings sheds no light on the causal parameter of interest. Moreover, for a given finite value of σ^2_B there comes a point at which observational data cease to be informative and where further advances to knowledge can come only from experimental findings. The illusion of observational learning is typically obscured by the way in which researchers conventionally report their nonexperimental statistical results. The standard errors associated with regression estimates, for example, are calculated based on the unstated but often implausible assumption that the bias associated with a given estimator is known with perfect certainty before the estimates are generated. These standard errors would grow much larger were they to take into account the value of a^2_B.

The only way to extract additional information from observational research is to obtain extrinsic information about its bias. By extrinsic information, we mean information derived from inspection of the observational procedures, such as the measurement techniques, statistical methodology, and the like. *Extrinsic information does not include the results of the observational studies and comparisons to experimental results.* If all one knows about the bias is that experimental studies produced an estimate of 10 while observational studies produced an estimate of 5, one's posterior estimate of the mean will not be influenced at all by the observational results.

To visualize the irrelevance of observational data with unknown biases, consider a hypothetical regression model of the form

$$Y = a + bX + U,$$

where ***Y*** is the observed treatment effect across a range of studies, ***X*** is a dummy variable scored 0 if the study is experimental and 1 if it is observational, and U is an unobserved disturbance term. Suppose that we have non-informative priors about ***a*** and ***b.*** The regression estimate of ***a*** provides an unbiased estimate of the true treatment effect.[2] Similarly, the regression estimate of ***b*** provides an unbiased estimate of the observational bias. Regression of course generates the same estimates of ***a*** and ***b*** regardless of the order in which we observe the data points. Moreover, the estimate of ***a*** is unaffected by the presence of observational studies in our dataset. This regression model produces the same estimate of a as a model that discards the observational studies and simply estimates

$$Y = a + U.$$

This point warrants special emphasis, since it might appear that one could augment the value of observational research by running an observational pilot study, assessing its biases by comparison to an experimental pilot study, and then using the new, more precise posterior of σ^2_B as a prior for purposes of subsequent empirical inquiry. The flaw in this sequential approach is that conditional on seeing the initial round of experimental and observational results, the distributions of ***M*** and ***B*** become negatively correlated. To update one's priors recursively requires a different set of formulas from the ones presented above. After all is said and done, however, a recursive approach will lead to exactly the same set of posteriors. As demonstrated in the Appendix, the formulas above describe how priors over ***M*** and ***B*** change in light of ***all*** of the evidence that subsequently emerges, regardless of the sequence in which these studies become known to us.

Although there are important limits to what observational findings can tell us about the causal parameter M, the empirical results may greatly influence posterior uncertainty about bias, a^2%. This posterior distribution is a weighted sum of three quantities: the prior belief about the location of the difference between the observational finding and the expected true effect, and the observed gap between the experimental and observational results. If the researcher enters the research process with diffuse priors about ***M*** and ***B*** such that $\sigma^2\boldsymbol{m} = \sigma^2_B = \infty$, the weights *q1*

[2] We are grateful to Doug Rivers, who suggested this analogy.

and q2 will be zero, and the posterior will reflect only the observed discrepancy between experimental and observational results. In this instance, the posterior variance reduces to the simple quantity

$$\sigma^2_{B|x_E,x_O,\sigma^2_M=\sigma^2_B=\infty} = \sigma^2_{x_E} + \sigma^2_{x_O},$$

which is the familiar classical result concerning the variance of the difference between two independent estimates. Note that this result qualifies our earlier conclusion concerning the futility of gathering observational data when one's priors are uninformative. When analyzed in conjunction with experimental results, observational findings can help shed light on the biases associated with observational research. Of course, this comes as small consolation to those whose primary aim is to learn about the causal parameter M.

A numerical example

A simple numerical example may help fix ideas. Consider two studies of the effects of face-to-face canvassing on voter turnout in a particular election in a particular city. The observational study surveys citizens to assess whether they were contacted at home by political canvassers and uses regression analysis to examine whether reported contact predicts voting behavior, controlling for covariates such as political attitudes and demographic characteristics (Kramer 1970; Rosenstone and Hansen 1993). The key assumption of this study is that reported contact is statistically unrelated to unobserved causes of voting. This assumption would be violated if reported contact were an imprecise measure of actual contact or if political campaigns make a concerted effort to contact citizens with unusually high propensities to vote. The majority of published studies on the effects of voter mobilization use some version of this approach.

The experimental study of face-to-face canvassing randomly assigns citizens to treatment and control groups. Canvassers contact citizens in the treatment group; members of the control group are not contacted.[3] This type of fully randomized experiment dates back to Eldersveld (1956)

[3] As Gerber and Green (2000) explain, complications arise in this type of experiment when canvassing campaigns fail to make contact with those subjects assigned to the treatment group. This problem may be addressed statistically using instrumental variables regression, although as Heckman and Smith (1995) note, the external validity of this correction requires the assumption that the canvassing has the same effect on those who could be reached as it would have had among those the canvassing campaign failed to contact. This assumption may be tested by randomly varying the intensity of the canvassing effort. Below, we take up the question of how our conclusions change as we take into account the potential for bias in experimental research.

and currently enjoys something of a revival in political science (Gerber and Green 2000).

Suppose for the sake of illustration that your prior beliefs about Ai, the effect of canvassing on voter turnout, were centered at 10 with a variance of 25 (or, equivalently, a standard deviation of 5). These priors imply that you assign a probability of about 0.95 to the conjecture that M lies between 0 and 20. You also hold priors about the observational bias. You suspect that contact with canvassers is measured unreliably, which could produce an underestimate of M, but also that contact with canvassers is correlated with unmeasured causes of voting, which may produce an overestimate of M. Thus, your priors about the direction of bias are somewhat diffuse. Let us suppose that B is centered at 2 with a variance of 36 (standard deviation of 6). Finally, you confront the empirical results. The experimental study, based on approximately 1,200 subjects divided between treatment and control groups, produces an estimate of 12 with a standard deviation of 3. The observational study, based on a sample of 10,000 observations, produces an estimate of 16 with a standard deviation of 1.

With this information, we form the posterior mean and variance for *M:*

$$E(M|X_e = 12, X_o = 16) = p_1(10) + p_2(12) + p_3(16 - 2) = 11.85$$

$$\sigma^2_{M|x_e,x_o} = \frac{1}{\frac{1}{25} + \frac{1}{9} + \frac{1}{1+36}} = 5.61$$

$$p_1 = \frac{5.61}{25} = .23,\ p_2 = \frac{5.61}{9} = .62,\ p_3 = \frac{5.61}{36+1} = .15.$$

Notice that although the observational study has much less sampling variability than the experimental study, it is accorded much less weight. Indeed, the experimental study has four times as much influence on the posterior mean as the observational study. The observational study, corrected for bias, raises the posterior estimate of M_3 while the prior lowers it, resulting in a posterior estimate of 11.9, which is very close to the experimental result. The prior variance of 25 has become a posterior variance of 5.6, a reduction that is attributable primarily to the experimental evidence. Had the observational study contained 1,000,000 observations instead of 10,000, thereby decreasing its standard error from 1 to 0.1, the posterior variance would have dropped imperceptibly from 5.61 to 5.59, and the posterior mean would have changed only from 11.85 to

11.86. A massive investment in additional observational data produces negligible returns.

One interesting feature of this example is that the experimental and observational results are substantively rather similar; both suggest that canvassing "works." Sometimes when experimental results happen to coincide with observational findings, experimentation is chided for merely telling us what we already know (Morton 2002: 15), but this attitude stems from the illusion described above. The standard error associated with the N = 10,000 observational study would conventionally be reported as 1, when in fact its root mean squared error is 6.1. In this situation, what we "already know" from observational research is scarcely more than conjecture until confirmed by experimentation. The posterior variance is four times smaller than the prior variance primarily because the experimental results are so informative.

The empirical results also furnish information about the bias associated with the observational data. The posterior estimate of $\boldsymbol{B}$ is 4.1, as opposed to a prior of 2. Because the experimental results tell us a great deal about the biases of the observational study, the posterior variance of $\boldsymbol{B}$ is 6.3, a marked decline from the prior value of 36. In advance of seeing the data, our priors over $\boldsymbol{M}$ and $\boldsymbol{B}$ were uncorrelated; afterwards, the posterior correlation between $\boldsymbol{B}$ and $\boldsymbol{M}$ becomes —0.92. This posterior correlation is important to bear in mind in the event that subsequent evidence becomes available. The estimating equations presented above describe how uncorrelated priors over $\boldsymbol{M}$ and $\boldsymbol{B}$ change in light of all of the evidence that emerges subsequently, regardless of the order in which it emerges. Thus, if a second observational study of 10,000 observations were to appear, we would recalculate the results in this section on the basis of the cumulative N = 20,000 observational dataset.

Allocating resources to minimize the posterior variance of *M*

Above we considered the case in which a researcher revises prior beliefs after encountering findings from two literatures, one experimental and the other observational. Now we consider a somewhat different issue: how should this researcher allocate scarce resources between experimental and observational investigation?

Suppose the research budget is R. This budget is allocated to experimental and observational studies. The marginal price of each experimental observation is denoted π_e; the price of a non-experimental observation

is π_o. Let n_e be the size of the experimental study; n_o the size of the observational study. The budget is allocated to both types of research subject to the constraint $\pi_e n_e + \pi_o n_o = R$.

Let the variance[4] of the experimental study equal $\sigma^2 x_e = \sigma^2{}_e / a n d$ the variance of the observational study equal $\sigma^2{}_{Xo} = \sigma^2{}_o/n_o$. Using the results in Theorem 1, the aim is to allocate resources so as to minimize the posterior variance of ***M,*** subject to the budget constraint ***R.***

Theorem 2. The optimal allocation of a budget ***R,*** given prices π_e and π_o, disturbance variances $\sigma^2{}_e$ and σ_o, and variance of priors about observational bias $\sigma^2{}_B$, takes one of three forms depending on the values of the parameters:

Case 1. For $\sigma_o^2(\pi_o/\pi_e) \geq \sigma_e^2$, allocate $n_2 = R/\pi_e$ and $n_o = 0$.

Case 2. For $\sigma_o^2(\pi_o/\pi_e)\left(1 + \frac{R\sigma_B^2}{\pi_o\sigma_\delta^2}\right)^{\frac{1}{2}} \geq \sigma_e^2 \geq \sigma_o^2(\pi_o/\pi_e)$,

allocate $n_o^* = \left[\left(\frac{\pi_e\sigma_\varepsilon^2}{\pi_o\sigma_\delta^2}\right)^{1-2} - 1\right]\left(\frac{\sigma_o^2}{\sigma_B^2}\right)$ and $n_e = \frac{R-\pi_o n_o^*}{\pi_e}$

Case 3. For $\sigma_e^2 \geq \sigma_o^2(\pi_o/\pi_e)\left(1 + \frac{R\sigma_B^2}{\pi_o/\sigma_\delta^2}\right)$, allocate $n_o = R/\pi_o$ and $n_o = 0$.

The implications of the Research Allocation Theorem in many ways parallel our earlier results. When allocation decisions are made based on uninformative priors about the bias ($\sigma^2{}_B = \infty$), no resources are ever allocated to observational research. As budgets approach infinity, the fraction of resources allocated to experimental research approaches 1. When $\sigma^2{}_B$ is zero, resources will be allocated entirely to either experiments or observational studies, depending on relative prices and disturbance variances.

The most interesting case is the intermediate one, where finite budgets and moderate values of $\sigma^2{}_B$ dictate an apportioning of resources between the two types of studies. Here, possible price advantages of observational research are balanced against the risk of bias. Particularly attractive, therefore, are observational studies that are least susceptible to bias, such as those based on naturally occurring randomization of

[4] The notation used in this section has been selected for ease of exposition. When the population variance for the subjects in an experimental study equals *v,* and *n* experimental subjects are divided equally into a treatment and control group, the variance for the experiment is *4v/n.* To convert from these units to the notation used here, define π_e as the cost of adding four subjects.

the independent variable (Imbens, Rubin, and Sacerdote 2001), near-random assignment (McConahay, 1982), or assignment that supports a regression-discontinuity analysis (Cook and Campbell 1979: ch. 3).

Notice that the allocation decision does not depend on σ^2_m, that is, prior uncertainty about the true causal parameter. How much one knows about the research problem before gathering data is irrelevant to the question of how to allocate resources going forward. Experimentation need not be restricted, for example, to well-developed research programs.

One further implication deserves mention. When the price-adjusted disturbance variance in observational research is greater than the disturbance variance in experimental research (see Case 1), all of the resources are allocated to experimental research. Reduction in disturbance variance is sometimes achieved in highly controlled laboratory settings or through careful matching of observations prior to random assignment. Holding prices constant, the more complex the observational environment, the more attractive experimentation becomes.

The Research Allocation Theorem provides a coherent framework for understanding why researchers might wish to conduct experiments. Among the leading justifications for experimentation are (1) uncertainty about the biases associated with observational studies; (2) ample resources; (3) inexpensive access to experimental subjects; and (4) features of experimental design that limit disturbance variability. Conversely, budget constraints, the relative costliness of experimental research, and the relative precision of observational models constitute leading arguments in favor of observational research. It should be emphasized, however, that the case for observational research hinges on prior information about its biases.

Discussion

In this concluding section, we consider the implications of these two theorems for research practice. We consider in particular (i) the possibility of bias in experiments; (ii) the value of methodological inquiry; (iii) the conditions under which observational data support unbiased causal inference, and (iv) the value of theory.

What about bias in experiments? In the preceding analysis, we have characterized experiments as unbiased and observational studies as potentially biased. It is easy to conceive of situations where experimental results are potentially biased as well. The external validity of an experiment hinges on three 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, and whether the response measure used in the experiment corresponds to the variable of interest in the population. In the example mentioned above, a canvassing experiment was conducted in a given city at a given point in time, Door-to-door canvassing was conducted by certain precinct workers using a certain type of get-out-the-vote appeal. Voter turnout rates were calculated based on a particular source of information. Extrapolating to other times, places, and modes of canvassing introduces the possibility of bias.

A straightforward extension of the present analytic framework could be made to cases in which experiments are potentially biased. Delete the expressions related to the unbiased experiment, and replace them with a potentially biased empirical result akin to the observational study. The lesson to be drawn from this type of analysis parallels what we have presented here. Researchers should be partial to studies, such as field experiments, which raise the fewest concerns about bias. The smaller the inferential leap, the better.

This point has special importance for the distinction between laboratory experiments and field experiments. Although lab experiments are often less costly than field experiments, the inferential leap from the laboratory to the outside world increases the risk of bias. Experiments often involve convenience samples, contrived interventions, and response measures that do not directly correspond to the dependent variables of interest outside the lab. These are more serious drawbacks than the aforementioned problems with field experiments. Field experiments may be replicated in an effort to sample different types of interventions and the political contexts in which they occur, thereby reducing the uncertainty associated with generalizations beyond the data. Replication lends credibility to laboratory experiments as well, but so long as the outcome variable observed in the lab (e.g., stated vote intention) differs from the variable of interest (actual voter turnout rates), and so long as there is reason to suspect that laboratory results reflect the idiosyncrasies of an artificial environment, the possibility of bias remains acute.

Disciplines such as medicine, of course, make extensive and productive use of laboratory experimentation. In basic research, animals such as rats and monkeys are used as proxies for human beings. One might suspect that the idiosyncratic biology and social environments of human beings would render this type of animal research uninformative, but in fact the correspondence between results obtained based on animal models and those based on human beings turns out to be substantial. Even stronger is the empirical correspondence between laboratory results involving non-random samples of human subjects and outcomes that occur when

medical treatments are deployed in the outside world. As Achen points out, when experience shows that results may be generalized readily from a laboratory setting, "even a tiny randomized experiment may be better than a large uncontrolled experiment" (1986: 7). Our point is not that laboratory experiments are inherently flawed; rather, the external validity of laboratory studies is an empirical question, one that has been assessed extensively in medicine and scarcely at all in political science.

All things being equal, the external validity of field experimentation exceeds that of laboratory experimentation. However, the advantages of field experimentation in terms of external validity may be offset by threats to internal validity that arise when randomized interventions are carried out in naturalistic settings. Heckman and Smith (1995) note that the integrity of random assignment is sometimes compromised by those charged with administering treatments, who deliberately or unwittingly divert a treatment to those who seem most deserving. They note also the complications that arise when people assigned to the control group take it upon themselves to obtain the treatment from sources outside the experiment. To this list may be added other sources of bias, such as problems of spillover that occur when a treatment directed to a treatment group affects a nearby control group or the experimenter's failure to measure variations in the treatment that is actually administered.

Whether a given field experiment confronts these difficulties, of course, depends on the nature of the intervention and the circumstances in which the experiment is conducted. When these threats to unbiasedness do present themselves, valid inference may be rescued by means of statistical correctives that permit consistent parameter estimation. When treatments are shunted to those who were assigned to the control group, for example, the original random assignment provides a valid instrumental variable predicting which individuals in the treatment and control groups actually received the treatment. Despite the fact that some members of the control group were treated inadvertently, the causal parameters of interest may be obtained using instrumental variables regression so long as the assigned treatment group was more likely to receive the treatment than the assigned control group (Angrist, Imbens, and Rubin 1996).[5]

Although statistical correctives are often sufficient to mend the problems that afflict field experiments, certain potential biases can only be

[5] The interpretation of these estimates is straightforward if one assumes (as researchers working with observational data often do) that the treatment's effects are the same for all members of the population. Strictly speaking, the instrumental variables regression provides an estimate of the effect of the treatment on those in the treatment group who were treated (or, in the case of contamination of the control group, those treated at different rates in the treatment and control groups).

addressed by adjusting the experimental design. For example, Howell and Peterson's (2002) experiments gauge the effects of private school vouchers on student performance by assigning vouchers to a random subset of families that apply for them. This design arguably renders a conservative estimate of the effects of vouchers, inasmuch as the competitive threat of a private voucher program gives public schools in the area an incentive to work harder, thereby narrowing the performance gap between the treatment group that attends private schools and the control group that remains in public school.[6] In order to evaluate the systemic effects of vouchers, it would be useful to perform random assignment at the level of the school district rather than at the level of the individual. In this way, the analyst could ascertain whether the availability of vouchers improves academic performance in public schools. This result would provide extrinsic evidence about the bias associated with randomization at the individual level.

Field experiments of this sort are of course expensive. As we point out in our Research Allocation Theorem, however, the uncertainties associated with observational investigation may impel researchers to allocate resources to field experimentation in spite of these costs. Observational inquiry may involve representative samples of the population to which the causal generalization will be applied and a range of real-world interventions, but the knowledge it produces about causality is more tenuous. The abundance of observational research and relative paucity of field experimentation that one currently finds in social science-even in domains where field experimentation is feasible and ethically unencumbered-may reflect excessive optimism about what is known about the biases of observational research.

The value of methodological inquiry. Our analytic results underscore not only the importance of unbiased experimental research but also the value of basic methodological inquiry. To the extent that the biases of observational research can be calibrated through independent inquiry, the information content of observational research rises. For example, if through inspection of its sampling, measurement, or statistical procedures the bias associated with a given observational study could be identified more precisely (lowering (σ^2_B), the weight assigned to those observational findings would go up. The social sciences have accumulated enormous amounts of data, but like ancient texts composed in some inscrutable language, these data await the discovery of insights that will enable them to become informative.

[6] The magnitude of this bias is likely to be small in the case of the Howell and Peterson interventions, which affected only a small percentage of the total student-age population.

Unfortunately, the *ex ante* prediction of bias in the estimation of treatment effects has yet to emerge as an empirical research program in the social sciences. To be sure, methodological inquiry into the properties of statistical estimators abounds in the social sciences. Yet, we know of no study that assesses the degree to which methodological experts can anticipate the direction and magnitude of biases simply by inspecting the design and execution of observational social science research.[7] The reason that this literature must be empirically grounded is that observational data analysis often involves a variety of model specifications; the statistical attributes of the final estimator presented to the reader may be very different from the characteristics of the circuitous estimation procedure that culminated in a final set of estimates. One potential advantage of experimentation is that it imposes discipline on data analysis because the statistical model is implied by the experimental design. Since data-mining potentially occurs in experimental analysis as well; the relative-merits of experimentation remain an open research question. This type of basic empirical investigation would tell us whether, as a practical matter, the uncertainty associated with observational bias is so great that causal inference must of necessity rely on experimental evidence.

A model for this type of literature may be found in survey analysis. The estimation of population parameters by means of random sampling is analogous to the estimation of treatment effects by means of randomized experimentation. Convenience and other non-random samples are analogous to observational research. In fields where random sampling methods are difficult to implement (e.g., studies of illicit drug use), researchers make analytic assessments of the biases associated with various non-random sampling approaches and test these assessments by making empirical comparisons between alternative sampling methodologies. The upshot of this literature is that biased samples can be useful so long as one possesses a strong analytic understanding of how they are likely to be biased.

When is observational learning not illusory? Sometimes researchers can approximate this kind of methodological understanding by seizing upon propitious observational research opportunities. For example, students of public opinion have for decades charted movements in presidential popularity. As a result, they have a clear sense of how much presidential popularity would be expected to change over the course of a few

[7] In the field of medicine, where plentiful experimental and observational research makes cross-method comparisons possible, scholars currently have a limited ability to predict *ex post* the direction and magnitude of observational bias (cf. Heinsman and Shadish 1996; Shadish and Ragsdale 1996). We are aware of no study showing that medical researchers can predict the direction of biases *ex ante*.

days. Although the terrorist attacks of September 11, 2001 were not the product of random assignment, we may confidently infer that they produced a dramatic surge in presidential popularity from the fact that no other factors can plausibly account for a sudden change of this magnitude. Expressed in terms of the notation presented above, the size of the observational estimate (x_O) dwarfs the bias term (B) and the uncertainty associated with it (σ^2_B), leaving little doubt that the terrorist attacks set in motion a train of events that increased the president's popularity.

Notice that the foregoing example makes use of substantive assumptions in order to bolster the credibility of a causal inference. In this case, we stipulate the absence of other plausible explanations for the observed increase in popularity ratings. Were we able to do this routinely in the analysis of observational data, the problems of inferences would not be so daunting. Lest one wonder why humans were able to make so many useful discoveries prior to the advent of randomized experimentation, it should be noted that physical experimentation requires no explicit control group when the range of alternative explanations is so small. Those who strike flint and steel together to make fire may reasonably reject the null hypothesis of spontaneous combustion. When estimating the effects of flint and steel from a sequence of events culminating in fire, σ^2_B is fairly small; but in those instances where the causal inference problem is more uncertain because the range of competing explanations is larger, this observational approach breaks down. Pre-experimental scientists had enormous difficulty gauging the efficacy of medical interventions because the risk of biased inference is so much greater, given the many factors that plausibly affect health outcomes.

This analytic framework leaves us with an important research principle: In terms of mean-squared error, it may be better to study fewer observations, if those observations are chosen in ways that minimize bias. To see this principle at work, look back at our numerical example and imagine that our complete data set consisted of 11,200 observations, 1,200 of which were free from bias. If we harbor uninformative priors about the bias in the observational component of the data set, the optimal weighting of these observations involves placing zero weight on the 10,000 bias-prone observations and focusing exclusively on the 1,200 experimental observations. The implication for large-N studies of international relations or comparative politics is that it may be better to focus on a narrow but uncontaminated portion of a larger data set.

How can one identify the unbiased component of a larger data set? One way is to generate the data through unbiased procedures, such as random assignment. Another is to look for naturally occurring instances where

similar cases are confronted with different stimuli, as when defendants are assigned by lot to judges with varying levels of punitiveness. More tenuous but still defensible may be instances where the processes by which the independent variables are generated have no plausible link to unobserved factors that affect the dependent variable. For example, if imminent municipal elections cause local officials to put more police on the streets, and if the timing of municipal elections across a variety of jurisdictions has no plausible relationship to trends in crime rates, the municipal election cycle can be used as an instrumental variable by which to estimate the effects of policing on crime rates (Levitt 1997). More tenuous still, but still arguably more defensible than an indiscriminant canvass of all available cases, is an attempt to match observations on as many criteria as possible prior to the occurrence of an intervention. For certain applications, this is best done by means of a panel study in which observations are tracked before and after they each encounter an intervention, preferably an intervention that occurs at different points in time. These methods are not free from bias, but the care with which the comparisons are crafted reduces some of the uncertainty about bias, which makes the results more persuasive.

The value of theory. The more theoretical knowledge the researcher brings to bear when developing criteria for comparability, the smaller the σ^2_B and the more secure the causal inference. Thus, in addition to charting a new research program for methodologists, our analysis calls attention to an underappreciated role of theory in empirical research. The point extends beyond the problem of case selection and internal validity. Any causal generalization, even one based on a randomized study that takes place in a representative sample of field sites, relies to some degree on extrapolation. Every experiment has its own: set of idiosyncrasies, and we impose theoretical assumptions when we infer, for example, that the results from an experiment conducted on a Wednesday generalize readily to Thursdays. Just as methodological insight clarifies the nature of observational bias, theoretical insight reduces the uncertainty associated with extrapolation.

But where does theoretical insight come from? To the extent that what we are calling theories are testable empirical propositions, the answer is some combination of priors, observational data, and experimental findings. Theory development, then, depends critically on the stock of basic, carefully assessed empirical claims. Social scientists tend to look down on this type of science as narrow and uninspiring, grasping instead for weak tests of expansive and arresting propositions. Unlike natural scientists, therefore, social scientists have accumulated relatively few secure empirical premises from which to extrapolate. This deficiency is unfortunate,

because developing secure empirical premises speeds the rate at which learning occurs as new experimental and observational results become available.

The framework laid out here cannot adjudicate the issue of how a discipline should allocate its resources to research questions of varying substantive merit, but it does clarify the conditions under which a given research program is likely to make progress. Our ability to draw causal inferences from data depends on our prior knowledge about the biases of our procedures.

Appendix: Proof of the proposition that the sequence of the empirical results does not affect the analytic results presented above

We seek to prove the following: if instead of observing the single experimental value x_e and the single observational value x_O we instead observe the series of experimental and observational sample mean values $x_e^{(1)}, x_e^{(2)}, \ldots x_e^{(k_e)}$ and $x_o^{(1)}, x_o^{(2)}, \ldots x_o^{(k_o)}$ such that

$$\frac{\sum_{j=1}^{k_e} n_e^{(j)} n_x^{(j)}}{\sum_{j=1}^{k_e} n_e^{(j)}} = x_e \text{ and } \frac{\sum_{j=1}^{k_o} n_o^{(j)} n_o^{(j)}}{\sum_{j=1}^{k_o} n_o^{(j)}} = x_o,$$

where x_e and x_O are the overall experimental and observational mean values, then once ***all*** of the data have been observed, the joint posterior density of $\boldsymbol{M}$ and $\boldsymbol{B}$ given the individual sample results will be equivalent to what one would have seen if the two overall means x_e and x_O were revealed simultaneously.

The proof follows directly from the sufficiency of the sample mean as an estimator for the population mean for normal distributions. Straight-forward application of Bayes' rule shows that, conditional on observing $x_e^{(1)}, x_e^{(2)}, \ldots x_e^{(k_e)}$ and $x_O^{(1)}, x_O^{(2)}, \ldots x_O^{(Ro)}$ the joint density of *M and B* is given by

$$f_{M,B|x_e^{(1)},x_e^{(2)},\ldots x_e^{(k_e)},x_o^{(1)},x_o^{(2)},\ldots x_e^{(k_e)}}(m,b)$$

$$= \frac{f_M(m) \times f_M(b) \times \prod_{j=1}^{k_e} f x_e^{(j)} | M = m\left(x_e^{(j)}\right) \times \prod_{j=1}^{k_e} f x_e^{(j)} | M = m, B = b\left(x_e^{(j)}\right)}{\int_{m'=-\infty}^{\infty} \int_{b'=-\infty}^{\infty} f M(m') \times f B(b') \times \prod_{j=1}^{k_e} f x_e^{(j)} | M = m'\left(x_e^{(j)}\right) \times \prod_{j=1}^{k_e} f x_e^{(j)} | M = m', B = b'(x_e^{(j)} db' dm'}$$

For normal distributions (see Freund 1971: 263) and constants C_e and C_o that do not depend on m and b, the likelihoods can be factored as

$$\prod_{j=1}^{k_e} f x_e^{(j)} = C_e \text{ x } f x_e | M = m\ (x_e) \text{ and}$$

$$\prod_{j=1}^{k_o} f x_o | M = m, B = b\left(x_o^{(j)}\right) = C_o \text{ x } f x_o | M = m, B = b(x_o).$$

Consequently, upon substituting into the numerator and denominator of the posterior density shown above, the constants C_e and C_o cancel, leaving the expression below:

$$f_{M,B|X_e=x_e,X_o=x_o}(m,b)$$

$$= \frac{f_M(m) \text{ x } f X_e | M = m(x_e) \text{ x } f B(b) \text{x } f X_o | M = m, B = b(x_o)}{\int_{m'=-\infty}^{\infty} \int_{b'=-\infty}^{\infty} f M(m') \text{ x } f X_e | M = m'(x_e) \text{ x } f B(b') \text{ x } f X_o | M = m', B = b'(x_o) db' dm'}$$

This expression is the same as the one presented in the text, which shows that the order in which we observe the empirical results does not matter.

The implication of this proof is that one cannot squeeze additional information out of observational research by comparing initial observational estimates to experimental estimates, calibrating the bias of the observational studies, and then gathering additional bias-corrected observational data. This procedure, it turns out, produces the same estimates as simply aggregating all of the observational and experimental evidence and using the estimation methods described in the text.

REFERENCES

Achen, Christopher H. 1986. *The Statistical Analysis Of Quasi–Experiments.* Berkeley: University of California Press.

Angrist, Joshua A. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80(3): 313–36.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identifica–tion of Casual Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (June): 444–55.

Berube, Danton. 2002. "Random Variations in Federal Sentencing as an Instru–mental Variable." Unpublished ms, Yale University.

Box, George E. P. and G. C. Tiao. 1973. *Baysian Inference in Statistical Analysis.* Reading, MA: Addison–Wesley.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi–Experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin.

Cook, Thomas D. and Monique R. Payne. 2002. "Objecting to the Objec–tions to Using Random Assignment in Educational Research," in Frederick Mosteller and Robert Boruch (eds.), *Evidence Matters: Randomized Trials in Education Research*. Washington, DC: Brookings Institution Press.

Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50 (March): 154–65.

Freund, John E. 1971. *Mathematical Statistics*, 2nd ed. New York: Prentice–Hall.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94: 653–63.

Green, Donald P. and Alan S. Gerber. 2003. "Reclaiming the Experimental Tradition in Political Science," in *The State of the Discipline III*, Helen Milner and Ira Katznelson (eds.), Washington, DC: American Political Science Association.

Gosnell, Harold F. 1927. *Getting–Out– The–Vote: An Experiment in the Stimulation* of Voting. *Chicago: University of Chicago Press.*

Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives 9* (Spring): 85–110.

Howell, William G. and Paul E. Peterson. 2002. *The Education Gap: Vouchers and Urban Schools.* Washington, DC: Brookings Institution Press.

Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote. 2001. "Esti–mating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Winners." *American Economic Review* 91(4): 778–94.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experi–ment." *Quarterly Journal of Economics* 116: 607–54. –

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton: Princeton University Press.

Kramer, Gerald H. 1970. "The Effects of Precinct–Level Canvassing on Voting Behavior." *Public Opinion Quarterly* 34 (Winter): 560–72.

Levitt, Stephen D. 1997. "Using Electoral Cycles in police hiring to Estimate the Effect of Police on Crime." *American Economic Review* 87(3): .270–90.

McConahay, John B. 1982. "Self–interest versus Racial Attitudes as Correlates of Anti–Busing Attitudes in Louisville: Is it the Buses or the Blacks?" *Journal of Politics* 44(3): 2–720.

Morton, Rebecca. 2002. "EITM: Experimental Implications of Theoretical Models." *The Political Meihodologist* 10(2): 14–16.

Rosenstone, Steven J. and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America.* New York: Macmillan Publishing Company.

Sherman, Lawrence W. and Dennis P. Rogan. 1995. "Deterrent Effects of Police Raids on Crack Houses: A Randomized, Controlled Experiment." *Justice Quarterly* 12(4): 755–81.

Slemrod, J., M. Blumenthal, and C. Christian. 2001. "Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota." *Journal of Public Economics* 79(3): 455–83.
Weiss, Carol H. 2002. "What to Do until the Random Assigner Comes," in Frederick Mosteller and Robert Boruch (eds.), *Evidence Matters: Randomized Trials in Education Research*. Washington, DC: Brookings Institution Press.