The Downstream Benefits of Experimentation

Donald P. Green and Alan S. Gerber
Department of Political Science, Yale University,
124 Prospect Street,
New Haven, CT 06520-8301
e-mail: donald.green@yale.edu
e-mail: alan.gerber@yale.edu

The debate about the cost-effectiveness of randomized field experimentation ignores one of the most important potential uses of experimental data. This article defines and illustrates “downstream” experimental analysis—that is, analysis of the indirect effects of experimental interventions. We argue that downstream analysis may be as valuable as conventional analysis, perhaps even more so in the case of laboratory experimentation.

1 Introduction

Although the value of randomized interventions is widely recognized, critics of field experimentation contend that large-scale randomized studies often fail to generate sufficient knowledge to justify their considerable cost. This view has been articulated most forcefully by Heckman and Smith (1995), who contend that randomized studies of job training programs and other policy interventions fail to show convincingly whether these interventions work. Although much of their critique concerns the practical problems that prevent truly random assignment of subjects, their larger objection is that field experiments do little to advance our theoretical understanding of how broader causal processes operate. The results of any particular intervention may say more about the idiosyncrasies of that experiment than the operation of broad causal forces, and experiments are seldom designed in ways that enable researchers to understand the mechanisms by which cause is translated into effect. Thus, Heckman and Smith conclude that field experimentation provides a less efficient path to knowledge than does the analysis of nonexperimental data.

Although there is much to commend in Heckman and Smith’s trenchant cost-benefit analysis, neither they nor defenders of field experimentation take into account what we call *downstream benefits of experimentation.* Experiments introduce exogenous variation into outcome variables, and one may then study how these variables affect other outcomes in turn. *Downstream benefits* refer to knowledge acquired when one examines the indirect effects of a randomized experimental intervention. This article begins by explicating these distinctions and suggesting the potential value of downstream experimental analysis. We then qualify our advocacy of this approach with a discussion of several threats to valid inference. These qualifications notwithstanding, we argue that a fair accounting of the benefits of experimental methodology must include the potential value of downstream experimental analysis.

2 Primary versus Downstream Experimental Investigation

Let Z ∈ {0,1} be a random variable that indicates whether a subject is assigned to treatment or control groups. For concreteness, let us suppose that Z refers to whether a student is randomly invited to participate in a school voucher program (Howell and Peterson 2002). In order to allow for the possibility that only some of the students will accept this invitation, let *X* ∈ {0,1} indicate whether a student actually uses a voucher to pay for schooling. The probability that a student uses the voucher conditional on a random invitation may be expressed as:

$$Pr(X = 1 \mid Z = 1) = x_t = a + b, \tag{1}$$

$$Pr(X = 1 \mid \mathrm{Z} = 0) = x_c = a. \tag{2}$$

In Eq. (1), students assigned to the treatment group use the voucher with *a* + *b* probability. In Eq. (2), the probability is simply *a.* Typically, no one in the Z = 0 control group actually uses a voucher, in which case x_c is zero. However, by distinguishing Z and *X,* our model makes allowance for imperfections in experimental procedure. In field settings, experimental interventions sometimes fail to treat every subject in the Z = 1 group (cf. Imbens and Rubin 1997).

Let *Y* refer to the outcome variable, in this case, whether a student receives a high school diploma. The probability of receiving a diploma may be expressed for the treatment and control groups.

$$Pr(Y = 1 | Z = 1) = y_t = c + dx_t, \tag{3}$$

$$Pr(Y = 1 | Z = 0) = y_c = c + dx_c. \tag{4}$$

Notice that the parameter c appears in both equations, reflecting the fact that these two randomly assigned groups have the same expected probability of receiving a diploma, holding constant their use of vouchers.

Because Z is assigned at random, it serves as an instrumental variable in the estimation of *d,* the average effect of voucher use on educational attainment (Angrist et al. 1996). In this example, the instrumental variables estimator may be derived by noting that

$$y_t - y_c = dx, - dx_c, \tag{5}$$

from which a consistent estimator of *d* may be formed using sample values of the probabilities $x_t, x_c, y_t,$ and y_c:

$$\hat{d} = \frac{\hat{y}_t - \hat{y}_c}{\hat{x}_t - \hat{x}_c}. \tag{6}$$

Thus, to estimate the marginal effect of using vouchers on the probability of receiving a high school diploma, take the observed difference between treatment graduation rates and control graduation rates and divide them by the observed difference between treatment voucher use and control voucher use. We would term this type of X-affects-y study a *primary* or *direct experimental analysis.*

What one makes of the parameter estimates that emerge from a primary experiment depends on the nature of the treatment and the subjects who participate in the study. School vouchers may work in some places and not others, depending on the socioeconomic environment, the nature of the school system, the student body, and so on. Moreover, there is no guarantee that the treatment *(X)* in one's experimental investigation will resemble the voucher program that would actually be implemented subsequently by policymakers. Given these uncertainties, experiments must be performed repeatedly in order to convey an understanding of the conditions under which treatment effects are large or small. Heckman and Smith argue against allocating resources to field experiments due to the number of studies necessary to isolate these contingent effects, and they instead recommend investing these resources in observational research.

Even if the idiosyncratic nature of specific interventions leaves us uncertain about their direct effects, their indirect effects remain theoretically informative. Suppose that our primary research interest were not the effects of a particular policy intervention, but rather the effects of characteristics that may vary across individuals. For concreteness, suppose we were eager to study the effects of educational attainment *(Y)* on voting (V). Scholars have long observed a strong correlation between voter turnout and educational attainment in U.S. survey data (e.g., Wolfinger and Rosenstone 1980). Although often interpreted to mean that each additional year of schooling produces an increase in one's propensity to vote, this statistical relationship could simply be the result of unobserved heterogeneity in nonexperimental data. Unobserved factors that cause people to obtain a high school diploma may also cause them to vote at higher rates. If so, regression of voter turnout on educational attainment produces biased estimates.

This inference problem can be circumvented by examining the effects of an exogenous change in educational attainment, brought about by random assignment of school vouchers. Let us express voter turnout as a function of school vouchers for the treatment and control groups:

$$Pr(V = 1 \mid Z = 1) \;—v_t = e + fy_t, \tag{7}$$

$$Pr(V = 1 \mid Z = 0) = v_c = e + fy_c. \tag{8}$$

Again, the initial random assignment of vouchers (Z) serves as an instrumental variable. Along the same lines as Eq. (6), a consistent estimator of f may be formed using sample estimates of $v_t, v_c, y_t,$ and y_c:

$$\hat{f} = \frac{\hat{v}_t - \hat{v}_c}{\hat{y}_t - \hat{y}_c}. \tag{9}$$

This estimator is consistent even when unobserved factors influence both educational attainment and voting (Gerber and Green 2001, pp. 83-84).

In this instance, the voucher experiment is valuable not only because it tells us about the role that vouchers may play in encouraging educational attainment, it also provides researchers the wherewithal to answer a largely unrelated question about how educational attainment affects voting behavior. Indeed, once an exogenous shock to education has been produced, one can investigate a range of hypotheses about the consequences of education. Does increased educational attainment reduce racial prejudice? Promote marital stability? Increase savings? Reduce smoking?

Appreciation of the downstream benefits of experimentation casts the role of experimental interventions in a different light. Those seeking to study the effects of education on voter turnout might well conduct randomized interventions that alter educational attainment. The logic of this exercise parallels the investigation of “natural experiments” (Rosenzweig and Wolpin 2000), in which scholars search for near-random processes in the environment (e.g., age cutoffs for entrance into elementary school) that generate more or less exogenous variation in the independent variable of interest (years of education completed). The difference between planned and natural experiments lies in the degree to which planned experimentation affords more precise control over the process of random assignment.

3 Caveats

Although the downstream analysis of experimental data has promise, it also confronts a series of practical and theoretical limitations. Researchers should bear in mind these limitations when they design and analyze experiments in this way.

3.1 *Statistical Power*

Examining the variance of the instrumental variables estimator presented in Eq. (9) makes apparent the fact that downstream experimentation is feasible only under certain circumstances. The estimator of *f* has an asymptotic variance of

$$VAR(f) = \frac{VAR(v_t - v_c) + f^2 VAR(y_t {}_- v_c) - 2fCOV(y_t - y_c, v_t - v_c)}{(y_t - y_c)^2}. \quad (10)$$

If the primary experiment reveals no effect, the denominator of this equation is zero, and the estimator has infinite variance. This result parallels the standard result that instrumental variables estimation is feasible only when the excluded instruments predict the independent variable of interest. In the example used here, the effects of educational attainment on voting can be estimated only when educational attainment is altered by the vouchers intervention. “Failed” experiments have no downstream utility. Conversely, the more profound the effects of the intervention, the greater the opportunity to study downstream consequences.

Another intuitive conclusion to be drawn from Eq. (10) is that downstream analysis is best performed on experiments that precisely reveal primary effects. The less uncertainty associated with experimental differences in educational attainment, the less uncertainty about the effects of educational attainment on voter turnout. Large *N* studies that examine the effects of profound interventions present the most fruitful opportunities for downstream investigation.

3.2 *Publication Bias*

From a practical standpoint, Eq. (10) implies that researchers may pick and choose among existing experiments, selecting those with large treatment effects for follow-up investigations. One must be cautious, however, when trolling through published experimental studies in search of downstream opportunities. Let us return to Eqs. (3) and (4), modifying them to allow for different intercepts:

$$y_t - c_t + dx_t, \quad (3)'$$

$$y_c = c_c + dx_c. \quad (4)'$$

Here, the treatment and control groups possess unobserved attributes that give them different probabilities of graduating, controlling for voucher use. For example, it may be that students in the treatment are more likely to come from affluent homes. The estimator in Eq. (6) tends to overestimate *d* when $c_t > c_c$, presenting the data analyst with what would seem to be a propitious opportunity for downstream research. However, consider what happens when a downstream analysis is performed on a data set for which $c_t > c_c$. Revised Eqs. (7) and (8) now allow for the possibility that the intercepts may differ for treatment and control groups. Imagine that the affluence of the treatment group makes them more likely to vote, controlling for educational attainment. In this case, the intercepts in (7) and (8) become a function of the values of c:

$$v_t =(e + qc_t) + fy_t \tag{7'}$$

$$v_c = (e + qc_c) + fy_c. \tag{8'}$$

When $c_t > c_c$ and $q > 0$, the estimator in Eq. (9) is no longer consistent. In other words, in the presence of unobserved heterogeneity—whereby unmeasured causes of education are related to unmeasured causes of voting—publication bias leads to biased downstream estimates. This point seems especially important to bear in mind in cases in which *d* is, in fact, zero, but sampling error generates an estimate of *d* that is significantly larger than zero. As Bound et al. (1995) explain in their discussion of weak instrumental variables, the estimator in Eq. (9) has the same expected value as a naïve regression of *V* on *Y*.

One must therefore be concerned about distortions as a result of publication bias. If the experiments that find their way into print represent a biased sample of all of the experiments that were actually performed because journals refuse to publish studies that report statistically insignificant estimates, a disproportionate number of experiments will report significant estimates of *d* when, in fact, *d* is zero. When *X* is assigned in ways that overstate both the treatment versus control differences in *Y* and V, downstream effect estimates are misleading. The practical implication of this point is that downstream researchers should be suspicious of experiments that involve a small number of subjects because it is here that publication biases are most severe (Gerber et al. 2001). Instead, downstream analysis should focus on larger and frequently replicated experiments.

3.3 *Direct Effects*

Equations (7) and (8) make the important assumption that the intercept parameters (e) are the same for treatment and control groups. As shown in the discussion of publication bias, a lot hinges on this assumption. Another way in which it may be violated occurs when the experimental intervention affects the downstream dependent variable directly. In the previous example, it seems unlikely that receiving an education voucher would directly alter one's propensity to vote. The concern arises in other instances, particularly those in which a given form of behavior is tracked over time. For example, in their study of voter mobilization, Gerber and Green (2000) randomly assigned registered voters to receive get-out-the-vote messages by phone, mail, or face-to-face canvass prior to the 1998 elections. Finding that the mail and face-to-face get-out-the-vote campaign significantly increased voter turnout in 1998, they sought to examine the effects of voting in 1998 on voting in 1999. If voting in 1998 did nothing to increase the likelihood of voting in 1999, one would expect to see the mail and face-to-face treatment and control groups vote at the same rates in 1999. However, ifvoting is habitforming, the contrast between treatment and control groups should persist to the next election. In fact, Gerber et al. (2000) found that the treatment group voted at a significantly higher rate in 1999, suggesting that voting in 1998 led to higher subsequent rates of voting.

The validity of this inference hinges on whether one can assume that the mobilization effort in 1998 had no direct influence on voting a year later. Perhaps voters resonated to the get-out-the-vote appeal 13 months after receiving mail, phone calls, or a face-to-face visit. This interpretation seems highly unlikely in light of the fact that campaigns deliberately hold back their contacts with voters until the closing months of an election campaign for fear that voters will forget their appeals if too much time elapses before election day. Still, it is important for researchers considering downstream consequences to grapple with the possibility of contamination through direct effects, addressing these concerns through additional experimentation. In this case, for example, concerns about direct effects could be ameliorated by an experiment demonstrating that mobilization efforts conducted several months prior to an election have no effect on voter turnout.

3.4 *Multiple Causes*

Closely related to the problem of direct effects is the question of what to do when a given intervention has multiple consequences. Imagine that the vouchers intervention not only affected educational attainment, as in Eqs. (3) and (4), but also political awareness, which we denote Y_I. Now consider what happens when both Y and Y_I affect the downstream probability of voting:

$$Pr(V = 1 \mid Z = 1) = v_t = e + fy_t + gy_{1t}, \quad (11)$$

$$Pr(V = 1 \mid Z = 0) = v_c = e + fy_c + gy_{1c}. \quad (12)$$

It is apparent that when g is nonzero, the estimator in Eq. (9) no longer provides consistent estimates of f. This estimation problem is not easily circumvented, even when Y_1 is observed. For example, one cannot treat Y_1 as an exogenous variable in an instrumental variables regression in which (Z, Y_1) are instruments and (Y, Y_1) are regressors; this estimator is inconsistent when Y_1 is correlated with unobserved causes of V.

In the presence of unobserved heterogeneity, consistent estimates of f and g can be obtained from an augmented experiment in which there are two randomized assignments (Z, Z_1), each of which leads to changes in Y and Y_I. Now (Z, Z_1) serve as instrumental variables in an estimator analogous to Eq. (9). The practical implication here is that more variegated experimental interventions make for more flexible and persuasive downstream analyses. By the same token, the most compelling downstream analyses are those based on primary experiments that produce a single outcome of relevance to V.

3.5 *Average Treatment Effects*

As researchers extrapolate from their experimental results to the outside world, they inevitably argue that their subjects and interventions are similar to those found in other settings. Many of the controversies surrounding experimentation focus on these issues of external validity. Concerns about external validity impinge on downstream investigations somewhat differently. In downstream research, the aim is not to achieve verisimilitude between an experimental intervention and an intervention that occurs in the real world. Rather, the aim is to draw secure inferences about the consequences of an exogenous shock to an independent variable of interest. Before one may generalize from a downstream study of school vouchers, one must ask whether voucher-induced educational attainment has different effects from educational attainment produced by some other causal mechanism.

The answer to this question naturally depends on the subject matter. The case of educational attainment is interesting in this regard. Rosenzweig and Wolpin (2000) argue that in some cases, interventions that bolster educational attainment do so disproportionately among people with lower levels of educational aptitude (because those with higher levels of aptitude are more likely to reach the upper limit of their potential regardless of these interventions). If so, the reported treatment effect is best understood as an average treatment effect among those with below average ability. When interpreting the downstream experimental results, the same kind of reasoning applies. A difference in voting rates between the treatment and control groups in the school vouchers experiment would tell us the effects of educational attainment for those people whose educational attainment tends to be altered by the offer of a voucher. Whether education's effects are distinctive in this population remains an open empirical question. One may look for interaction effects within this population in order to ascertain the degree to which f varies, or one may replicate the experiment using other populations in which f is expected to be different.

4 Laboratory Experiments from a Downstream Perspective

The difficulty of moving from experimental results to the outside world is of special concern for laboratory experiments in the social sciences due to the artificial nature of the lab and the fact that subjects are often drawn from the ranks of college undergraduates, whose outlook on the world may defy extrapolation to other populations. Even those studies that make use of nonstudents and attempt to create naturalistic environments confront problems of external validity. Studies of the effects of television news on political opinions, for example, typically invite subjects into a simulated living room, expose them to doctored news programs, and elicit their post-treatment views by means of an opinion survey. When such studies report randomly induced opinion change, it is unclear what to conclude about the influence of TV news outside the lab. Is this intervention comparable to what typically occurs in naturalistic settings? What is the relationship between the dependent variable measured in the lab (e.g., stated vote intention) and its real-world analog (e.g., voter turnout)? At this point, the discussion of these experiments often descends into a squabble between those who believe in the external validity of these experiments and those who do not.

Downstream experimentation puts these kinds of laboratory experiments to new use, focusing on the information they provide about the indirect consequences of opinion change. Any opinion that can be altered through randomized media exposure is potentially an independent variable in a subsequent analysis. Bear in mind that public opinion researchers routinely treat political opinions as causative forces: beliefs about the economy are said to influence the popularity of the incumbent president, which in turn is said to alter the extent to which citizens identify with the president's political party, which in turn affects vote choice. These inferences grow out of decades of nonexperimental research, and the question is whether they would be supported by evidence derived from downstream experimentation. So long as a laboratory intervention can be assumed to exert no direct influence over the downstream dependent variables, the experiment may provide valuable information.

Concerns about external validity do not vanish in the context of downstream analysis, of course. Downstream results obtained in laboratory settings warrant replication elsewhere in order to ensure that the results are not particular to the nature of the primary experimental intervention. Still, one senses that external validity does not pose the same challenge to downstream experiments that they do for primary experiments. The question is not whether a given piece of TV news coverage changes subjects' beliefs about the state of the economy, but whether these newly altered economic assessments change their political views. By shifting the focus away from the particular manner in which changes in beliefs are introduced, downstream analysis removes one of the main impediments to external validity.

However, concerns about heterogeneous treatment effects remain. Those with little interest in public affairs may be more susceptible to new information about the economy and more prone to change their political views in the wake of this information. For this reason, downstream analyses based on laboratory experiments must attend to the possibility of individual differences, a concern that again underscores the importance of replication using different types of subjects.

5 Conclusion

To date, the merits of randomized experimentation have been discussed almost exclusively from the standpoint of what we have termed *primary experimental analysis*. We would argue that the analysis of indirect experimental effects warrants attention in its own right. Indeed, in cases such as laboratory experimentation, the scientific value of the research may lie principally in what it can tell us about downstream effects. Social scientists should therefore take notice when the independent variables of interest to them are the dependent variables in another scholar's experiment, particularly if the intervention is known to have sizeable effects. Where it is feasible to return to the experimental subjects and gather data on downstream dependent variables, the opportunities for fruitful research are considerable.

The interdisciplinary implications of the previous point bear emphasis. Social scientists would be well advised to take stock of experimentation in other disciplines and, indeed, to participate in the design and execution of these studies. Unlike most calls for interdisciplinary collaboration, this one does not rest on the conviction that the social sciences should become more theoretically or methodologically ecumenical. The intellectual forces that impel sociologists, psychologists, political scientists, and economists to conduct randomized experiments are, in some sense, irrelevant. Any randomized intervention that genuinely produces different outcomes in treatment and control conditions is potentially useful for reasons that may have little to do with the hypotheses that originally inspired the experiment. Random perturbations of household income induced by an economist's experiment may be of interest to the psychologist studying affluence and levels of happiness, the sociologist studying income and health behaviors, and the political scientist studying attitudes toward redistribution.

As we point out, drawing secure inferences from these downstream experiments is not unproblematic. Experimental analysis requires the investigator to impose certain statistical assumptions about how the initial treatment relates to other variables. Yet as is true for any methodological argument, the question is how downstream inference compares to the available alternatives. The preponderance of empirical research in social science involves either observational studies or laboratory experimentation, each of which presents formidable barriers to valid inference and generalization. At a minimum, downstream experimental investigation presents an untapped opportunity to supplement and cross-validate observational and laboratory studies.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Casual Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444-455.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables Is Weak." *Journal of the American Statistical Association* 90:443-450.

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* 94:653-663.

Gerber, Alan S., and Donald P. Green. 2001. "Do Phone Calls Increase Voter Turnout? A Field Experiment." *Public Opinion Quarterly* 65:75-85.

Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. "Testing for Publication Bias in Political Science." *Political Analysis* 9:385-392.

Gerber, Alan S., Donald P. Green, and Roni Shachar. 2000. "Voting May Be Habit Forming." Unpublished manuscript.

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

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

Imbens, GuidoW., and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25:305-327.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38:827-874.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.