

Experimental reasoning in social science¹

Andrew Gelman²

7 Aug 2010

As a statistician, I was trained to think of randomized experimentation as representing the gold standard of knowledge in the social sciences, and, despite having seen occasional arguments to the contrary, I still hold that view, expressed pithily by Box, Hunter, and Hunter (1978) that “To find out what happens when you change something, it is necessary to change it.”³

At the same time, in my capacity as a social scientist, I’ve published many applied research papers, almost none of which have used experimental data.⁴

In the present article, I’ll address the following questions:

1. Why do I agree with the consensus characterization of randomized experimentation as a gold standard?
2. Given point 1 above, why does almost all my research use observational data?

In confronting these issues, we must consider some general issues in the strategy of social science research. We also take from the psychology methods literature a more nuanced perspective that considers several different aspects of research design and goes beyond the simple division into randomized experiments, observational studies, and formal theory.

My practical advice is that we should be doing more field experiments but that simple comparisons and regressions are, and should be, here to stay. We can always interpret such analyses descriptively, and description is an important part of social science, both in its own right and in providing foundations upon which to build formal models. Observational studies can also be interpreted causally when attached to assumptions which can be attacked or defended on their own terms.

Beyond this, the best statistical methods for experiments and observational studies are not so different. There are historical and even statistical reasons why experimentalists have focused on

¹ For *Field Experiments and their Critics*, ed. Dawn Teele. Yale University Press.

² Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, <http://www.stat.columbia.edu/gelman/> We thank Dawn Teele for soliciting this article and the Institute for Education Sciences for grant R305D090006-09A.

³ Box, Hunter, and Hunter refer to “changing” something—that is experimentation—without reference to randomization, a point to which we return later in this article.

⁴ I will restrict my discussion to social science examples. Social scientists are often tempted to illustrate their ideas with examples from medical research. When it comes to medicine, though, we are, with rare exceptions, at best ignorant laypersons (in my case, not even reaching that level), and it is my impression that by reaching for medical analogies we are implicitly trying to borrow some of the scientific and cultural authority of that field for our own purposes. Evidence-based medicine is the subject of a large literature of its own (see, for example, Lau, Ioannidis, and Schmid, 1998).

simple differences whereas observational researchers use regression, matching, multilevel models, and other complex tools. But, as Susan Stokes points out in her article in this volume, interactions are crucial in understanding social science research. And once you start looking at interactions, you're led inexorably to regression-type models and complex data summaries.

There is an analogous issue in survey research. Real-world surveys have complicated patterns of nonavailability and nonresponse and are typically adjusted using poststratification on demographic variables such as age, sex, ethnicity, education, and region of the country (see, for example, Voss, Gelman, and King, 1995). These adjustments can be performed via multilevel regression and poststratification (Gelman and Little, 1997), in which a statistical model such as logistic regression is fit, conditional on all variables used in the nonresponse adjustment, and then the estimates from this model are averaged over the general population—poststratified—using census information or other estimates of the distribution of the demographic variables.

If you had a simple random sample of Americans, say, you wouldn't need to do poststratification if your only goal is to estimate the national average. But why stop at that? Once you realize you can estimate public opinion by state and by subgroup within each state (Lax and Phillips, 2009), you'll want to fit the full model even if your data happened to come from a random sample. The statistical adjustment tools originally developed for handling nonresponse turn out to be useful even when nonresponse is not an issue.

To return to causal research: the issues of inference and design are important—and that is the subject of most of the present volume—but the statistical modeling issues are *not* so strongly affected. Whether your data are observational or experimental, you'll want to construct a model conditional on variables that can affect treatment assignment and also variables that can have potentially important interactions—and, in practice, these two sets of variables often overlap a lot. This is a point implicitly made by Greenland, 2005 (and by Imai, King, and Stuart, 2008 in the social science context): the most important sources of bias commonly arise from treatment interactions which are interesting in their own right.

The same sort of analysis that researchers use to adjust for potential confounders in an observational study is also useful for accounting for interactions in an experimental context. True, with experimental data, you *can* get clean estimates of average causal effects using simple averages. But why do that? Field experiments are often expensive and have small sample sizes; why not get the most out of your data? And here I'm not just talking about estimating higher moments of a distribution but rather estimating nonlinear relationships and interactions. From this statistician's perspective, it's a crime to spend a million dollars on data collection and then do a five-dollar statistical analysis.

Starting from the goal, or starting from what we know

Policy analysis (and, more generally, social science) proceeds in two ways. From one direction, there are questions whose answers we seek—how can we reduce poverty, fight crime, help people live happier and healthier lives, increase the efficiency of government, better translate public preferences into policy, and so forth? From another direction, we can gather discrete bits of understanding about pieces of the puzzle: estimates of the effects of particular programs as

implemented in particular places. A large part of social science involves performing individual studies to learn useful bits of information, and another important aspect of social science is the synthesis of these “stylized facts” into inferences about larger questions of understanding and policy. Much of my own social science work has gone into trying to discover and quantify some stylized facts about American politics, with various indirect policy implications which we generally leave to others to explore.

Much of the discussion in the present volume is centered on the first of the approaches described in the paragraph above, with the question framed as: What is the best design for estimating a particular causal effect or causal structure of interest? Typical quantitative research, though, goes in the other direction, giving us estimates of the effects of incumbency on elections, or the effects of some distribution plan on attitudes, or the effect of a particular intervention on political behavior, and so forth. Randomized experiments are the gold standard for this second kind of study, but additional model-based quantitative analysis (as well as historical understanding, theory, and qualitative analysis) is needed to get to the larger questions.

It would be tempting to split the difference in the present debate and say something like the following: Randomized experiments give you accurate estimates of things you don’t care about; Observational studies give biased estimates of things that actually matter. The difficulty with this formulation is that inferences from observational studies also have to be extrapolated to correspond to the ultimate policy goals. Observational studies can be applied in many more settings than experiments but they address the same sort of specific micro-questions. For all the reasons given by Gerber, Green, and Kaplan, I think experiments really are a better choice when we can do them, and I applaud the expansion in recent years of field experiments in a wide variety of areas in political science, economics, sociology, and psychology.

I recommend we learn some lessons from the experience of educational researchers, who have been running large experiments for decades and realize that, first, experiments give you a degree of confidence that you can rarely get from an observational analysis; and, second, that the mapping from any research finding—experimental or observational—is in effect an ongoing conversation among models, data, and analysis.

My immediate practical message here is that, before considering larger implications, it can be useful to think of the direct and specific implications of any study. This is clear for simple studies—any estimated treatment effect can also be considered as a descriptive finding, the difference between averages in treatment and control groups, among items that are otherwise similar (as defined by the protocols of the study).

Direct summaries are also possible for more complicated designs. Consider, for example, the Levitt (1977) study of policing and crime rates, which can be viewed as an estimate of the causal effect of police on crime, using political cycles as an instrument, or, more directly, as an estimate of the different outcomes that flow from the political cycle. Levitt found that, in cities with mayoral election years, the number of police on the street goes up (compared to comparable city-years without election) and the crime rate goes down. To me, this descriptive summary gives me a sense of how the findings might generalize to other potential interventions. In particular, there

is always some political pressure to keep crime rates down, so the question might arise as to how one might translate that pressure into putting police on the street even in non-election years.

For another example, historical evidence reveals that when the death penalty has been implemented in the United States, crime rates have typically gone down. Studies have found this at the national and the state levels. However, it is difficult to confidently attribute such declines to the death penalty itself, as capital punishment is typically implemented in conjunction with other crime-fighting measures such as increased police presence and longer prison sentences (Donohue and Wolfers, 2006).

In many ways I find it helpful to focus on descriptive data summaries, which can reveal the limits of unstated model assumptions. Much of the discussion of the death penalty in the popular press as well as in the scholarly literature (not in the Donohue and Wolfers paper, but elsewhere) seems to go to the incentives of potential murderers. But the death penalty also affects the incentives of judges, juries, prosecutors, and so forth. One of the arguments in favor of the death penalty is that it sends a message that the justice system is serious about prosecuting murders. This message is sent to the population at large, I think, not just to deter potential murderers but to make clear that the system works. Conversely, one argument against the death penalty is that it motivates prosecutors to go after innocent people, and to hide or deny exculpatory evidence. Lots of incentives out there. One of the advantages of thinking like a “statistician”—looking at what the data say—is that it gives you more flexibility later to think like a “social scientist” and consider the big picture. With a narrow focus on causal inference, you can lose this.

Research designs

I welcome the present exchange on the pluses and minuses of social science experimentation but I worry that the discussion is focused on a limited perspective on the possibilities of statistical design and analysis.

In particular, I am concerned that “experiment” is taken to be synonymous with “randomized experiment.” Here are some well-known designs which have some aspect of randomization or experimentation without being full randomized experiments:

Natural experiments. From the Vietnam draft lottery to zillions of regression discontinuity studies, we have many examples where a treatment was assigned not by a researcher—thus, not an “experiment” under the usual statistical definition of the term—but by some rule-based process that can be mathematically modeled.

Non-randomized experiments. From the other direction, if a researcher assigns treatments deterministically, it is still an experiment even if not randomized. What is relevant in the subsequent statistical analyses are the factors influencing the selection (for a Bayesian treatment of this issue, see Gelman et al., 2003, chapter 7).

Sampling. Textbook presentations often imply that the goal of causal inference is to learn about the units who happen to be in the study. Invariably, though, these are a sample from a larger population of interest. Even when the study appears to include the entire population—for

example, an analysis of all 50 states—the ultimate questions apply to a superpopulation such as these same states in future years.

Mixed experimental-observational designs. In practice, many designs include observational variation within an experiment. For example, Barnard et al. (2003) analyze a so-called “broken experiment” on school vouchers. Mixing of methods also arises when generalizing experimental results to new decision problems, as we discuss in the meta-analysis example later in this article.

Informal experimentation. One point often made by proponents of causal inference from observational data is that people (and, for that matter, political actors) do informal experimentation all the time.

Our brains can do causal inference, so why can't social scientists? Humans do (model-based) everyday causal inference all the time (every day, as it were), and we rarely use experimental data, certainly not the double-blind stuff that is considered the gold standard. I have some sympathy but some skepticism with this argument (see Gelman et al., 2010), in that the sorts of inferences used as examples by the proponents of “everyday causal reasoning” look much less precise than the sorts of inferences we demand in science (or even social science).

In any case, our everyday causal reasoning is *not* purely observational. As Sloman (2005) points out, one of the purposes of informal experimentation in our ordinary lives is to resolve some of the causal questions left open by models and observational inference. In our lives, we experiment all the time, on matters as small as trying out new recipes or new routes to work, to trying out new teaching methods in our classes, to the big personal decisions such as cohabitation as a potential prelude to marriage.

Formal experimentation. What are the minimal conditions a study must have to be counted as an “experiment” in the statistical sense. Informal experimentation, as described above, is not enough. In my view, any experiment has two defining features:

1. *Deliberately assigned treatments.* Simply trying a new idea is not an experiment unless you make a clear decision of when and where you will try it and its alternatives. Without this clarity, you risk having an observational study in which conditions are assigned endogenously and then only labeled as experimental treatments after the fact.
2. *Systematic data collection,* which includes a measurement of the experimental treatments (what, exactly, was done in each case) and of the outcome, and also, ideally, of pre-treatment variables. Without a measurement protocol there is no way to estimate treatment effects with any validity.

These are, under my definition, the only *necessary* features of a formal experiment. But of course we would like to have much more. Techniques such as randomization, blindness, large sample size, and measurement of background variables (to better allow extrapolation to the general population) allow us to have much more confidence in our inferences. As usual in statistics, the less care that goes into the data collection, the more that inferences are sensitive to model assumptions.

Self-experimentation. One way to focus on experimentation—in isolation of related but distinct ideas such as randomization—is to consider the most basic form, where the sample size is 1 and the experimenter and the subject are the same. Self-experimentation has a long history in medical research (Altman, 1986) and more recently it has been advocated as a research method to be used more widely.

Seth Roberts is a professor of psychology, with a background in rat learning experiments who has used self-experimentation to generate and study hypotheses about sleep, mood, and nutrition. Here are some of his findings (Roberts, 2004): “Seeing faces in the morning on television decreased mood in the evening and improved mood the next day . . . Standing 8 hours per day reduced early awakening and made sleep more restorative . . . Drinking unflavored fructose water caused a large weight loss that has lasted more than 1 year . . .” Self-experimentation generates new hypotheses and is also an inexpensive way to test and modify them, with the sort of flexibility that might be difficult to attain in an NIH-funded study of experimental volunteers.

These conditions are similar to those found in social science research and policy analysis, especially at the national level where it is difficult to go beyond $n=1$. Practical policymaking is, in many ways, a form of self-experimentation on the local, state, or national level. Looked at from this perspective, an important research step is to go beyond informal trying-out of ideas toward formal self-experimentation with its clearly-defined treatments and measurement protocols. Successful small experiments—randomized or not—can lay the foundation for larger, more conclusive studies, although there are challenges involved in taking this last step (see Gelman and Roberts, 2007).

Our love for the full field-experimentation package (including randomization, blindness, and large sample sizes) should not blind us to the advantages of experimentation in its simplest form. Good self-experimentation includes manipulations (that is, experimentation) but also careful and dense measurements—“self-surveillance.” Similarly, designers of observational studies and field experiments alike should be aware of the benefits to be gained by extensive measurement and also by exploratory data analysis—those statistical tools that allow us to check our assumptions and generate new ideas as well as to estimate and test fixed, pre-chosen hypotheses.

Example: Interactions in a meta-analysis

I now return to the second question posed at the beginning of this article: Given the manifest virtues of experiments, why do I almost always analyze observational data? The short answer is almost all the data out there are observational.

Rather than give a list of dozens of research projects, I will discuss a particular example, not one of my most important projects but one in which, even though we were working with data from clean randomized experiments, our ultimate analysis was observational.

Several years ago, some colleagues and I were involved in a survey with a disappointingly low response rate. We did some research into how we could do better and discovered a paper by Singer et al. (1999) with a meta-analysis on the use of incentives to increase participation in surveys. Each experiment included in the meta-analysis had been conducted by randomly

assigning different incentive conditions to participants in a survey. (Survey participants are already randomly selected and so it is nearly effortless to embed an experiment within.) There were between two and five different conditions in each survey-experiment, with a total of 101 conditions in 39 surveys. Each condition had several descriptors (the dollar value of the incentive, whether the incentive was offered before or after the interview, whether the survey was conducted face to face or by telephone, whether the incentive was in cash or a gift, and the burden of the survey (whether it was long or short) and an outcome—the response rate of the people interviewed under that survey condition.

Singer et al. ran a regression of the increases in response rate (compared to the no-incentive condition) for these surveys and estimated an effect of 1.4 percentage points, plus .34 percentage points for each additional dollar of incentive. There was also some evidence that incentives were more effective when given before the interview (using an earlier mail contact). These estimates all made sense but we did not fully believe some of the other results of the model. For example, the estimated effect for gift versus cash incentive was very large in the context of the other effects: the expected effect of a postpaid cash incentive of \$10 in a low-burden survey was $1.4 + 10 \cdot .34 - 6.9 = -2.1\%$, thus actually lowering the response rate.

We find it implausible that giving a gift would lower the response rate. But the estimate is based on a randomized experiment, so what grounds do we have to distrust it? The answer is that, although each individual study in the meta-analysis is experimental, the comparison of conditions *among* studies is observational.

Because of the nonrandomized design (which is unavoidable because the 39 different studies were conducted at different times with different goals), coefficient estimates cannot automatically be given direct causal interpretations, even if they are statistically significant. The estimated effect of -6.9% in response rate for a gift (compared with the equivalent incentive in cash) is presumably an artifact of interactions in the data between the form of the incentive and other variables that affect response rates. To put it most simply, the surveys in which gifts were considered as an experimental condition may be surveys in which, for some other reasons, incentives happened to be less effective.

The only consistently randomized factor in the experiments is the incentive indicator itself; the other factors are either observational (burden, mode) or experimental but generally not assigned randomly (value, timing, form). This is a common problem when a meta-analysis is used to estimate a “response surface” rather than simply an average effect (see Rubin, 1989).

In this particular analysis, the implausible estimate from treatment interactions that had not been included in the model. We addressed the problem in the usual fashion for observational data, by fitting a regression model including interactions of the effect of incentive with survey, along with some interactions of the predictive factors with each other (Gelman, Chan, and Stevens, 2003). Finally, having fit a model that we found plausible (albeit, with some of the coefficients being less than two or even one standard error away from zero), we applied it to the conditions of the survey on which we were working. For this application, high levels of interactions are a modeling necessity, not merely a theoretical possibility, as we were considering various options for our survey.

Conclusions

In social science as in science in general, formal experiments (treatment assignment plus measurement) teach us things that we could never observe from passive observation or informal experimentation. I applaud the increasing spread of field experiments and recommend that modern statistical methods be used in their design and analysis. It is an inefficiency we cannot afford, and which shows insufficient respect for the participants in our surveys and experiments, to use the simplest statistical methods just because we can. Even in the unlikely setting that treatments have been assigned randomly according to plan and that there are no measurement problems, there is no need to limit ourselves to simple comparisons and estimates of average treatment effects.

In areas of design, measurement, and analysis, field experimenters can learn much from researchers in sample surveys (for the problem of extending from sample to population which is often brought up as a concern with experiments) and from researchers in observational studies (for the problem of modeling complex interactions and response surfaces). And observational researchers—that would be most empirical social scientists, including me—should try our best to model biases and to connect our work to solid experimental research wherever possible.

References

Altman, L. K. (1986). *Who Goes First: The Story of Self-Experimentation in Medicine*. Berkeley: University of California Press.

Barnard, J., Frangakis, C., E., Hill, J. L, and Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association* 98, 299-323.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. New York: Wiley.

Dehejia, R. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? A postscript. <http://www.nber.org/papers/w11442/postscript.pdf>

Dehejia, R., and Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053-1062.

Donohue, J. J., and Wolfers, J. (2006). Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review* 58, 791-845.

Gelman, A. (2010). Causality and statistical learning. *American Journal of Sociology*, to appear.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: Chapman and Hall.

- Gelman, A., and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* 23, 127-135.
- Gelman, A., and Roberts, S. (2007). Weight loss, self-experimentation, and web trials: A conversation. *Chance* 20 (4), 57-61.
- Gelman, A., Stevens, M., and Chan, V. (2003). Regression models for decision making: A cost-benefit analysis of incentives in telephone surveys. *Journal of Business and Economic Statistics* 21, 213-225.
- Gerber, A. S., Green, D. P., and Kaplan, E. H. (2010). The illusion of learning from observational research. In the present volume.
- Greenland, S. (2005). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society A* 168, 267-306.
- Imai, K., King, G., and Stuart, E. A. (2010). Misunderstandings between experimentalists and observationalists about causal inference. In the present volume.
- Lau, J., Ioannidis, J. P. A., and Schmid, C. H. (1998). Summing up evidence: one answer is not always enough. *Lancet* 351, 123-127.
- Lax, J., and Phillips, J. (2009). How should we estimate public opinion in the states? *American Journal of Political Science* 53.
- Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87, 270-90.
- Roberts, S. (2004). Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight (with discussion). *Behavioral and Brain Sciences* 27, 227-288.
- Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer-Verlag.
- Rubin, D. B. (1989). A new perspective on meta-analysis. In *The Future of Meta-Analysis*, eds. K. W. Wachter and M. L. Straf. New York: Russell Sage Foundation.
- Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999). The effects of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics* 15.
- Slooman, S. (2005). *Causal Models: How People Think About the World and Its Alternatives*. Oxford University Press.
- Smith, J., and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? (with discussion). *Journal of Econometrics* 120, 305-375.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* 33, 230-251.

Stokes, S. C. (2010). A defense of observational research. In the present volume.

Voss, D. S., Gelman, A., and King, G. (1995). Pre-election survey methodology: details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly* 59, 98-132.