Ch. 4: Statistical research designs for causal inference[*]

Fabrizio Gilardi[†]

January 24, 2012

1 Introduction

In Chapter 3 we have discussed the different ways in which the social sciences conceptualize causation and we have argued that there is no single way in which causal relationships can be defined and analyzed empirically. In this chapter, we focus on a specific set of approaches to constructing research designs for causal analysis, namely, one based on the potential outcomes framework developed in statistics. As discussed in Chapter 3, this perspective is both probabilistic and counterfactual. It is probabilistic because it does not assume that the presence of a given cause leads invariably to a given effect, while it is counterfactual because it involves the comparison of actual configurations with hypothetical alternatives that are not observed in reality. In essence, this approach underscores the necessity to rely on comparable groups in order to achieve valid causal inferences. An important implication is that the *design* of a study is of paramount importance. The way in which the data are produced is the critical step of the research, while the actual data analysis, while obviously important, plays a secondary role. However, a convincing design requires research questions to be broken down to manageable pieces. Thus, the big tradeoff in this perspective is between reliable inferences on very specific causal relationships on the one hand, and their broader context and complexity (and, possibly, theoretical relevance) on the other hand.

The chapter first distinguishes between two general perspectives on causality, namely, one that puts the *causes* of effects in the foreground, and another that is more interested in the *effects* of causes. We will then introduce the potential outcomes framework before discussing

[*] Chapter for Maggetti, Martino, Fabrizio Gilardi and Claudio M. Radaelli (2012), *Research Design in the Social Sciences*, SAGE Publications.

[†] Associate Professor, Department of Political Science, University of Zurich, Switzerland. Email: gilardi@ipz.uzh.ch; URL: http://www.fabriziogilardi.org/.

several research designs for causal inference, notably various types of experiments and quasi-experiments. The last section discusses the implications for research design, and the conclusion summarizes the main points.

2 Causes of effects and effects of causes

To understand the specificities of statistical research designs for causal inference, it is useful to consider a general difference between quantitative and qualitative approaches to causal analysis. While the former typically focus on the "effects of causes," the latter usually examine the "causes of effects" (Mahoney and Goertz, 2006). An equivalent distinction is that between "forward" and "reverse" causal inference: the former asks "What might happen if we do X?" while the second asks "What causes Y?" (Gelman, 2011). The difference between the two approaches overlaps in part with that characterizing "variable-oriented research" on the one hand, and "case-oriented research" on the other (Ragin 1987, 34–68; see also Chapter 3). Obviously, both are legitimate and fruitful perspectives in the social sciences, each with their own tradeoffs.

For instance, consider the case of women's quotas in parliamentary elections. Figure 1 compares the percent of women in parliament in 69 countries with and 84 countries without quotas (Tripp and Kang, 2008). Each dot represents a country, and Finland, Sweden, France, and the Netherlands are highlighted. Horizontal lines represent the average percent of women in parliament in each group. From an effects-of-causes perspective, we would investigate the consequences of quotas on female representation. That is, the starting point is the presumed cause (quotas), and the goal is measuring its causal connection with the presumed effect (e.g., the percent of women in parliament). The fact that, on average, countries with quotas have more women in parliament than those without quotas suggests that quotas might be conducive to better female representation. On the other hand, in a causes-of-effects perspective we would begin with the outcome and trace our way back to the possible causes. For instance, we could ask why two relatively similar countries such as Finland and the Netherlands have similar shares of women in parliament (about 37%), although only the latter has gender quotas. We could also ask why, in Sweden, there are almost four times as many women in parliament as in France (45.3% vs. 12.2%), given that both countries have introduced quotas. The first perspective would likely produce a single estimate of the causal effect, while the second would probably give

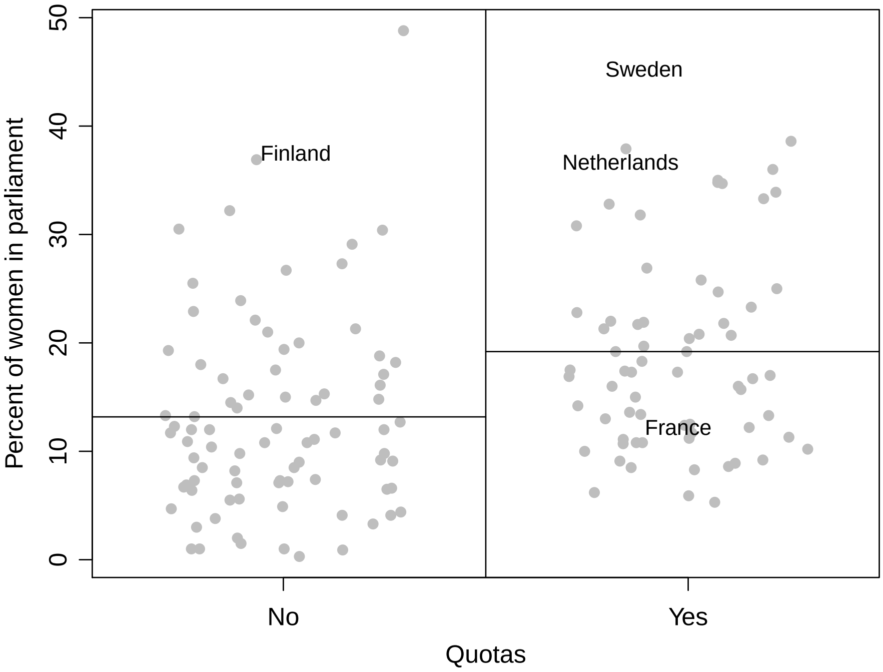

Figure 1: *Percent of women in parliament in 69 countries with and 84 countries without quotas. Each dot represents a country. Horizontal lines represent the average percent of women in parliament in each group (own elaboration based on Tripp and Kang 2008).*

an extensive account of the numerous factors influencing female representation and explain the cases holistically, that is, in all their complexity.

Statistical research designs embrace the first approach. As Gelman (2011) argues, "What causes Y?" is often the question that motivates the analysis in the first place. However, attempting to answer the question directly leads inevitably to a proliferation of hypotheses, most of which are actually likely to have *some* validity. Thus, the risk is that the analysis becomes intractable. This is the problem of overdetermination, or the fact that there are always a myriad of factors contributing in some way to a specific outcome. As we will discuss in Chapter 6, there are methods that allow to address this issue from a case-oriented perspective, that is, within a causes-of-effects approach. However, statistical designs reframe the question in terms of the effects of causes. They break the question down, identify a particularly interesting factor, and ask what consequences it has on the outcome of interest. An implication of this strategy is that multiple analyses are needed to uncover complex causal paths, because each analysis can only examine one at a time. Or, as Gelman (2011) puts it, in this perspective we are trying to learn about a specific causal path within a more complex causal structure, but not about the causal structure itself. Thus, statistical designs prioritize the reliability of very specific causal estimates at the expenses of the broader context in which they operate and possibly even of the connection with the original (theoretical and/or empirical) problem, which must be redefined in order to make it fit within the strict requirements of the analytical design.

3 The potential outcomes framework

The potential outcomes framework, also known as the counterfactual model, presupposes a dichotomous treatment (D_i), such as (to continue our example from the previous section) the presence or absence of women's quotas. If $D_i = 1$, it means that country i has quotas for the representation of women in parliament, while if $D_i = 0$, it means that it does not. Further, the framework assumes that there are two potential outcomes for each unit i, Y_{1i} and Y_{0i}. The outcomes are associated with the two possible values of the treatment. In our example, Y_{1i} is the percent of women in parliament in country i in the presence of quotas, while Y_{0i} is that percent if the same country i does not have quotas. Formally, we can represent this idea as follows:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Notice that both outcomes refer to the same unit. But, of course, it is impossible that, in our example, the same country both does and does not have quotas. This is why the two outcomes are called "potential:" only one is realized and can be observed, while the other is its logical counterpart, which exists only in the realm of ideas. However, conceptually, both are necessary for the definition of a causal effect. If we were able to observe, for the same country i, the percent of women both with and without quotas, then we could compute the causal effect for that country simply as the difference between the two outcomes:

$$Y_{1i} - Y_{0i}.$$

On this basis, and always assuming that both outcomes can be observed (which, in fact, is not possible), we can define two other quantities. The first is the average treatment effect (ATE), which, as the name indicates, is the average effect of the treatment for all units (for instance, the average effect of quotas in all countries):

$$\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_{1i} - Y_{0i}).$$

That is, the ATE is defined as the average difference between the two potential outcomes in all countries. The second quantity is the average treatment effect on the treated (ATT or ATET), that is, the effect of the treatment averaged only over units that actually receive the treatment (for instance, the average effect of quotas in countries with quotas):

$$\text{ATT} = \frac{1}{\sum_{i=1}^{n} D_i}\sum_{i=1}^{n} D_i(Y_{1i} - Y_{0i}).$$

That is, we make the same computation as for the ATT, but only for the subset of countries with quotas (i.e., for which $D_i = 1$). Countries without quotas ($D_i = 0$) are disregarded.

These definitions rely on a critical assumption, namely, the so-called stable unit treatment value assumption (SUTVA) (Morgan and Winship, 2007, 37–40). It has two components. First, the treatment must be the same for all units. While the *effect* of the treatment can vary across units (if it did not, we would not need to compute averages for the ATE and ATT), the treatment itself must be equivalent in all units. In our example, this assumption is in fact violated because there are several types of quotas, namely, compulsory or voluntary party quotas, reserved lists, and women-only lists (Tripp and Kang, 2008, 347). By collapsing them in a simple "quotas vs no quotas" dichotomy, we assume that each of these instruments has the same consequences for female representation, which is unlikely to be the case. However, this assumption is necessary in the potential outcomes framework. Second, the outcomes in one unit must be independent of the treatment status in other units. In other words, the percent of women in a given country must be unrelated to whether or not other countries have quotas. This assumption should be met in our example, but it easy to imagine other situations in which it does not hold, for instance, when the treatment has network effects or other types of externalities. The interdependencies discussed in Chapter 7 are good cases in point.

As we have already said, these definitions of treatment effects are purely theoretical. In reality, we cannot observe the same unit both with and without the treatment. This is known

Table 1: *The fundamental problem of causal inference (based on Morgan and Winship, 2007, 35).*

	(% women if quotas) Y_{1i}	(% women if no quotas) Y_{0i}
(Quotas) $D_i = 1$	$Y_{1i} \vert D_i = 1$ (Observable)	$Y_{0i} \vert D_i = 1$ (Counterfactual)
(No quotas) $D_i = 0$	$Y_{1i} \vert D_i = 0$ (Counterfactual)	$Y_{0i} \vert D_i = 0$ (Observable)

as the "fundamental problem of causal inference" (Holland, 1986), and it is what makes causal inference so difficult in practice. The nature of the problem is summarized in Table 1. In reality we can observe two outcomes, namely, in our example, the percent of women in parliament in the presence of quotas given that there are actually quotas, and that percent in the absence of quotas given that there are actually no quotas. However, to compute the quantities defined above, we would need also the two corresponding counterfactual outcomes, namely, the percent of women in parliament in the absence of quotas in countries that actually have quotas, and that percent in the presence of quotas in countries that actually have quotas. To illustrate more intuitively, take the case of France. Because this country has women's quotas, we are here in the top-left corner of Table 1. To compute the causal effect of quotas in France, we should take the difference between the observed percent of women in parliament (12.2%) and the value that we would observe if France had no quotas—that is, the value of the top-right corner of Table 1. The same logic applies to countries that have no quotas, namely, those in the bottom-right corner, which would need to be compared with their counterfactuals in the bottom-left corner.

What if we compute the difference between the two quantities we can actually observe? As we have seen in Figure 1, countries with quotas have, on average, more women in parliament (19.2%) than those without them (13.2%). It turns out that this observed difference in averages is equal to the ATT (one of our quantities of interest), plus a selection bias (Angrist and Pischke, 2009, 14). In our example, the selection bias corresponds to the average difference between the percent of women in parliament without quotas in countries that actually have quotas (a counterfactual) and the percent without quotas in countries that actually do not have them (which is observable). The former group includes countries such as France, Germany, and

Table 2: *Countries with and without quotas are quite different (own calculations based on Tripp and Kang, 2008).*

	Quotas	No quotas	Difference	Sig. level
GDP/cap (log)	7.96	7.11	0.85	0.001
Women's education	47.82	46.22	1.61	0.096
Democracy	2.55	2.04	0.52	0.000
Electoral system	0.57	0.29	0.28	0.000

Sweden, while the latter includes countries such as Ghana, Syria, and Vietnam. In fact, Table 2 shows that the two groups differ systematically in a number of ways. Countries with quotas tend to be wealthier, more democratic, and more likely to have a proportional system. Although the difference is only borderline significant, women in countries with quotas also tend to be more educated. All these factors are likely to be associated with a higher share of women in parliament even in the absence of quotas. This is what "selection bias" means in this context. Countries are not assigned quotas randomly; they self-select into this policy. Therefore, countries with and without quotas differ in a number of ways and the two groups are not well comparable.

In sum, within the potential outcomes framework, causal effects are clearly defined but cannot be directly computed in practice because the required counterfactuals are unobservable. However, researchers can rely on several methods to estimate them. We turn to them in the next section.

4 Methods

4.1 Regression

In most quantitative studies, the default research design is regression analysis with observational data, that is, information that was not generated by a process controlled by the researcher. The dataset used by Tripp and Kang (2008) is a typical example. By contrast, experimental data are those produced under the supervision of the researcher. Continuing with our example, a bivariate regression of the share of women in parliament on quotas indicates that countries with quotas have on average about 6% more women in parliament than countries without quotas, and that the difference is statistically highly significant.[1] This difference corresponds exactly to

[1] % women = 13.18 (1.03) + 6.02 (1.53) × quotas. OLS estimates, standard errors in parentheses.

what is shown in Figure 1. An obvious problem with this analysis is that it fails to control for the differences that exist across countries beyond the presence of quotas, such as those shown in Table 2. In other words, the bivariate regression neglects the selection bias problem. A multivariate regression can mitigate it, to a certain extent. If we include the variables listed in Table 2, quotas remain significantly associated with female representation, but the size of the effect is reduced by half in comparison with the bivariate regression. That is, controlling for GDP per capita, women's education, democracy, and the type of electoral system, countries with quotas have on average only about 3.2% more women in parliament than countries without quotas.[2] The inclusion of control variables is known also as "covariate adjustment," which means that the analysis adjusts the estimate of the causal effect for those covariates (i.e. variables) that can be taken into account.

Under some conditions, regression can yield unbiased estimates of causal effects (Morgan and Winship, 2007, 136–142). These conditions, however, are quite restrictive and generally unlikely to be met in practice.

First, there must be no omitted variables in the analysis. That is, in our example, all factors influencing the percent of women in parliament besides quotas must be measured and included in the regression. Obviously, no analysis can ever fulfil this requirement perfectly, which means that only rarely can the causal estimates produced by regression analysis be credibly considered unbiased.

Second, the functional relationship between the control variables and the outcome must be fully and correctly specified. This means, for instance, that any nonlinearities in the relationship between say, GDP per capita and women's representation, as well as any interactions (for instance, between GDP per capita and women's education) must be explicitly and correctly modeled. This becomes quickly intractable with even just a handful of variables, a problem that is known as the "curse of dimensionality." This requirement stems from the fact that, in most practical situations, the treatment and control groups are quite different or, in other words, the covariates are not balanced between them. In fact, this is the case in our example, as shown in Table 2. Therefore, the analysis needs to make assumptions in order to extrapolate the comparison between countries with and without quotas for specific combinations of control

[2] % women = -1.67 (5.68) + 3.2 (1.55) × quotas + 6.02 × electoral system − 0.11 (1.16) democracy + 0.11 (0.14) × women's education + 1.18 (0.59) × GDP/cap (log). OLS estimates, standard errors in parentheses.

variables. The problem can be alleviated by a method called "matching" (Ho et al., 2007), which attempts to make the treated and control groups more similar by removing "incomparable" cases. Concretely, one can for instance compute the probability that a unit receives the treatment (the "propensity score") and then find, for each treated unit, an untreated unit with a very similar propensity score. If the procedure is successful (which depends on the characteristics of the dataset), then a better balance between the two groups is achieved (that is, they are more comparable) and the analysis becomes less dependent on the specific assumptions made by the regression model. However, matching improves comparability only with respect to variables that can actually be observed. Thus, the first condition (no omitted variables) remains a big problem.

4.2 Experiments

As we have seen, there are two main practical problems when implementing the potential outcomes approach empirically. First, selection bias is ubiquitous, which means that the comparability of the treatment and control groups is usually limited. Second, while regression can in principle solve this problem, omitted variables and the "curse of dimensionality" will in most cases lead to biased estimates of causal effects. The appeal of the experimental approach is that it is much more effective in ensuring that treated and control units are in fact comparable. This occurs through "randomization," namely, random assignment of treatment to the units. Specifically, what defines experiments is that randomization is done by researchers themselves. With randomization, systematic differences between the two groups can occur only by chance and, if the number of units is sufficiently large, with a very low probability. Moreover, the procedure works for both observable and unobservable characteristics, such that omitted variables are no longer a problem. Because randomization is so powerful, the data can in principle be analyzed with simple techniques and the difference in means for the outcome between treatment and control groups (or, equivalently, the coefficient of a bivariate regression) can be interpreted as the ATE as well as the ATT. A common problem is that units are not selected randomly from the population, such that it is not possible to generalize the estimates straightforwardly beyond the sample. However, the estimates are still valid for the units that were part of the experiment. It should be emphasized that, of course, randomization is not magic and there are several ways in which it can go wrong. For instance, it is possible that not all the units that are assigned to the treatment are actually treated or, conversely, that some control units become exposed to it ("noncompliance"); it is also possible that, for one reason or another, outcomes cannot be observed for some units ("attrition") (Gerber and Green, 2008). However, experiments have an unparalleled capacity to uncover causal relationships and are widely considered the "gold standard" in this respect.

In our women's quotas example, an experiment would imply that quotas are attributed to countries randomly. As a consequence, and in contrasts to what we have seen in Table 2, the groups of countries with and without quotas would be very similar, if not exactly identical, on all characteristics that could potentially affect women's representation, including those that cannot be observed. Therefore, the average difference in the percent of women in parliament between the two groups could in principle be interpreted as the causal effect of quotas. The example shows the advantages of the experimental approach, but also an obvious drawback in the social sciences. In many, if not most, cases, randomization cannot be implemented for a number of practical and ethical reasons. Thus, it is not surprising that experiments are not the first method that comes to mind when thinking of social science research. At the same time, in recent years they have been used with increasing frequency and success and they have become a mainstream tool for social scientists (Druckman et al., 2006). We can distinguish among three broad types, namely, laboratory, survey, and field experiments, which we discuss in the next sections.

4.2.1 Laboratory experiments

Laboratory experiments are "experiments where the subjects are recruited to a common location, the experiment is largely conducted at that location, and the researcher controls almost all aspects in that location, *except* for subjects' behavior" (Morton and Williams, 2008, 346, original emphasis). They are what first comes to mind when we hear the word "experiment," namely, a relatively small group of people, not necessarily representative of the broader population (e.g., students), following precise instructions to perform set of abstract tasks.

Despite their stylized nature, laboratory experiments can help to uncover important causal relationships. For example, Correll (2004) was interested in how cultural beliefs about gender differences in ability affect career choices through the self-assessment of performance. If it is commonly accepted in society that, say, men are better than women at math, then the theory is that, at equal levels of objective skills, men will evaluate their competence higher than women do. Consequently, men will be more inclined than women to pursue a career in a field where math is important, thus reproducing existing gender imbalances. To estimate the causal effect of cultural frames, Correll (2004) set up an experiment in which about 80 undergraduate students were asked to perform a test purportedly designed to develop a new examination for graduate school admission. The test had no right or wrong answers (but was perceived as credible) and all subjects were given the same score, that is, the same objective assessment of their skills. By contrast, their cultural expectations (namely, the treatment) were manipulated by assigning subjects randomly to two groups. The treated group was told that males tend to perform better at the task, while the control group was informed that there are usually no gender differences in this context. After completing the test and receiving the (fake) scores, subjects were asked to provide a self-assessment of their performance and to answer questions about how likely they would be to pursue a career requiring high levels of the skills that were purportedly tested. In line with the theoretical expectations, the analysis showed that, under the treatment condition, women's self-assessment was lower than males', and that males' assessment under the treatment was higher than under the control condition. Further, these biased self-assessment were related to potential career plans.

A second example is Dunning and Harrison (2010), which studied how cross-cutting cleavages moderate the political saliency of ethnicity. The theory is that ethnic differences play a more important role in politics if citizens speaking a given language, for instance, belong to a different religion and are poorer than those speaking other languages. If, however, the different cleavages (linguistic, religious, economic) are not superposed in this way, then it is expected that language is less relevant as a determinant of political behavior. Dunning and Harrison (2010) studied this argument in the case of Mali, a highly ethnically diverse country, by focusing on "cousinage," namely, a form of identity and social bonds connected with groups of patronymics (surnames) but distinct from ethnicity. The 824 subjects of the experiments, recruited in Mali's capital city, were shown videotaped political speeches by a purported political independent considering being a candidate for deputy in the National Assembly. Subjects were asked to evaluate the candidate on a number of dimensions. The treatment was the politician's last name, which subjects could readily associate with both ethnicity and cousinage ties. This setup yielded four combinations of subjects' and politician's ethnicity and cousinage, namely, same ethnicity/cousins, same ethnicity/not cousins, different ethnicity/cousins, and different ethnicity/not cousins. Additionally, in the control group the politician's name was not given. In line with theoretical expectations, the candidate was evaluated best by the subjects when they shared both ethnicity and cousinage and worst in the opposite scenario. Additionally, cousinage compensated ethnicity: the candidate was evaluated similarly when subjects and candidate were from the same ethnic group but without cousinage ties and when they were from a different ethnic group but with cousinage ties.

In order to produce valid results, laboratory experiment must consider an extensive list of potential problems, such as the nature of experimental manipulations, location, artificiality, subjects' selection and motivation, and ethical concerns (for a thorough discussion, see Morton and Williams, 2010). Furthermore, they are vulnerable to the objection that, while their internal validity may be strong (that is, their results are valid within the context of the experiment), their conclusions cannot be generalized to the "real world." We will return to this point in the conclusion.

4.2.2 Survey experiments

Survey experiments randomly assign the respondents of a survey to control and treatment conditions through the manipulation of the form or placement of questions (Gaines, Kuklinski and Quirk, 2007, 3–4). Because many survey experiments use samples that are representative of the population, they promise to achieve both internal and external validity, the first through randomization, and the second through representativeness (Barabas and Jerit, 2010, 226). These potential qualities, in combination with an increasingly easy and cheap access to survey resources, have made survey experiments more popular among social scientists in recent years.

For example, Hainmueller and Hiscox (2010) examined attitudes toward immigration. They asked whether, as predicted by the labor market competition model, people tend to oppose immigrants with a skills level similar to their own, who would be perceived as a more direct threat in the competition for jobs. The experiment was embedded in a survey completed by 1,601 respondents in the United States, who were randomly divided in two groups. Those in the treatment group were asked whether they agreed that the US should accept more highly skilled immigrants from other countries. The question asked in the control group was identical, but "highly skilled" was replaced with "low-skilled." The authors could confirm that randomization worked well because the distribution of respondents' characteristics in the two groups were statistically undistinguishable. The main finding of the analysis is that, contrary to theory, both low-skilled and highly skilled respondents prefer highly skilled immigrants, which suggests that noneconomic concerns are very important to explain attitudes toward immigration.

Another example is Linos (2011), who studied cross-national interdependencies (one of the topics of Chapter 7) in the field of family policy with an experiment in which 1,291 Americans were asked whether they agree that the United States should increase taxes to finance a paid maternity leave. Respondents were assigned randomly to a control group, in which the question was formulated neutrally, or to one of four treatment groups. In the first and second, respondents were informed that the proposed policy was already in place in France or in most Western countries, respectively. In the third, respondents learned that the policy was recommended by the United Nations. Finally, in the fourth the policy was endorsed by "American family policy experts." The results show that, while in the control group only 20% of respondents support increasing taxes to pay for maternity leave, the percent jumps to about 40% in the treatment groups referring to France or other Western countries. Interestingly, the effect of foreign models is comparable to that of American experts, while that of the UN is even slightly higher. Thus, foreign experiences seem to play a significant role in shaping public opinion on family policy, which could be an important channel through which policies spread cross-nationally.

Researchers employing survey experiments face a distinct set of issues (Gaines, Kuklinski and Quirk, 2007; Barabas and Jerit, 2010). The treatment can be problematic in several ways. It is typically administered as a single exposure to an artificially intense stimulus, while in reality people may be exposed to it at varying degrees, at several points in time, and in combination with other factors. Moreover, exposure to the real-world version of the treatment prior to the survey can bias the results. Also, survey experiments usually measure the immediate effects of the treatment, but it would be important to know how long they last. In short, even if the sample is representative, external validity can be compromised if the treatment itself lacks representativeness.

4.2.3 Field experiments

Field experiments "are experiments where the researcher's intervention takes place in an environment where the researcher has only limited control beyond the intervention conducted" (Morton and Williams, 2008, 346). The central characteristic of experiments (randomized treatment assignment) is preserved but takes place in the "real world," which complicates its implementation in various ways. Field experiments are well established in particular in the study of political behavior and the political economy of development, but they have caught on also in other subfields.

For instance, Olken (2010) studied a classic question of democratic theory, namely, the comparative advantages of direct democracy and representation. The field experiment randomized the political process through which infrastructure projects were selected in 49 Indonesian villages. About 53% of the villages were randomly assigned to a direct democratic process in which all adults eligible to vote in national elections could express their preference. In the remaining villages the standard process was followed. Project selection took place in small meetings open to the public but that, in fact, were attended by a limited number of members of the local elite (such as government officials and representatives of various groups). On average, about 20 times as many people participated in the referenda than in the meetings. The randomization produced treatment and control groups that were statistically indistinguishable with respect to both village characteristics (e.g., ethnic and religious fragmentation, distance to subdistrict capital, population) and individual characteristics (e.g., education, gender, age, occupation). The results of the experiment showed that the same projects were selected under both decision-making processes, which suggests that representation does not lead to outcomes that are biased in favor of the elite's preferences. However, villagers were significantly more satisfied with the decisions when they were taken through referenda. Thus, it seems that the main effect of direct democracy is to increase the legitimacy of decisions, but not necessarily to shift their content closer to the population's preferences.

Another field experiment attempted to uncover the effects of political advertising on voters' preferences by randomizing radio and television ads, for a total value of about $2 million, during the 2006 reelection campaign of Texas governor Rick Perry (Gerber et al., 2011). The study randomized both the starting date and the volume of ads across 20 media markets in Texas, but not stations or programs. The outcome, namely, voters' evaluation of the candidate, was measured using large daily polls. Results showed a strong short term effect of the ads. The maximum ads volume was was associated with an increase of almost 5 percent points in the candidate's vote share during the week in which the ads were aired. However, this effect vanished as soon as a week afterwards. Thus, the results suggest that political advertizing does make a difference, but this evaporates quite quickly.

In addition to problems common to all experiments (such as external validity), field experiments present some specific challenges (Humphreys and Weinstein, 2009, 373–376). Given that many interesting variables cannot be randomized because of practical constraints, only a relatively small subset of questions can be investigated with this method. A possible solution is to focus on smaller units (e.g., municipalities instead of countries), but this will reduce the external validity of the analysis. Because field experiments take place in real time and in real settings, there are many factors that are not under the control of researchers and that can therefore contaminate the findings. A common problem are spillovers, or the fact that intervention in one unit may affect outcomes in other units. As discussed in Section 3, this violates the SUTVA assumption of the potential outcomes framework. The logistics of field experiments also constrains their size and reduce the precision of the estimates, which is a problem especially if the size of the effects is small. Finally, because they operate in real contexts, field experiments also raise certain ethical concerns.

4.3 Quasi-experiments

Quasi-experiments are observational studies (that is, using data that were not generated by a process controlled by the researcher) in which, thanks to circumstances outside the researcher's control, random treatment assignment is approximated to a certain extent. That is, although the assignment of units to treatment or to control status is not determined by the researchers but by naturally occurring social and political processes, some features of the procedures make it credible to assume that is is "as if at random." As Dunning (2008) argues, the plausibility of this assumption is variable and the burden of proof must be on the researcher. Thus, it is useful to situate quasi-experiments on a continuum with standard observational studies at one end and classical randomized experiments at the other. Quasi-experiments can take different forms. We discuss three, namely natural experiments, discontinuity designs, and instrumental variables.

4.3.1 Natural experiments

In natural experiments, the "as if at random" component comes from some social, economic, and/or political process that separates two groups cleanly on a theoretically relevant dimension. That is, although the quasi-randomization occurs without the researcher's intervention, it produces well-defined treatment and control groups.

For instance, Hyde (2007) studied the effects of international election monitoring on electoral fraud with data from the 2003 presidential election in Armenia, using polling stations as units of analysis. The outcome variable was the share of votes of incumbent president Kocharian, who was widely believed to have orchestrated extensive fraud operations. Poll stations in the treatment group were those visited by international observers, while those in the control group were not inspected by the monitors. To measure the treatment status of poll stations, Hyde (2007) relied on the list of assigned polling stations produced by the organization in charge of monitoring the elections, the Office for Democratic Institutions and Human Rights of the Organization for Security and Co-operation in Europe (OSCE/ODIHR). The validity of the natural experiment rests upon the assumption that international observers were assigned to polling stations in a way that approximates random assignment, and Hyde (2007) discussed in detail why this assumption is plausible in this case. The OSCE/ODIHR staff completed the lists arbitrarily, only on the basis of logistical considerations and with no knowledge of the socio-economic and political characteristics of the polling stations. The analysis showed that the incumbent presidents received significantly more votes (about 2–4%) in stations that were not monitored in comparison with those that were visited by observers, which suggests that this control mechanism has an impact on the extent of electoral fraud.

In another study, Bhavnani (2009) exploited an actual randomization, albeit one which he did not design, to investigate the long-term effects of quotas on female representation, that is, their consequences after they are withdrawn. A policy initiative in India reserved a certain number of seats for women in local elections, which were were chosen randomly for one legislature. The goal of this selection procedure was not to allow an evaluation of the policy (though this was a welcome side-product), but rather to make it as fair as possible by ensuring that men would be excluded from certain seats only temporarily, and without biases towards specific seats. Reserved and unreserved seats were statistically indistinguishable on many relevant dimensions, which suggests that the randomization likely worked. The analysis of elections in 1997 and 2002 showed that not only had quotas an effect on female representation during the election in which they were enforced, which must be true if the policy is implemented properly. They had an effect also in the next election, that is, after they were no longer in force. Comparing districts that were open both in 1997 and in 2002 with those that were reserved in 1997 but open again in 2002, the percent of female winners was significantly higher in the latter (21.6% vs. 3.7%). This indicates that the effects of quotas extend beyond their duration, possibly by introducing new female candidates into politics and by changing the perceptions of voters and parties.

Natural experiments are appealing because they feature randomization in a real-world setting without the direct involvement of the researcher. However, because researchers have no control on them, and because good natural experiments are rare, they often originate from the availability of a convenient configuration instead from a previously defined research question. In this sense, they tend to be method-driven rathe than problem-driven. Nonetheless, this is not necessarily problematic and the examples that we have just seen prove that natural experiments can be used to investigate important questions.

4.3.2 Discontinuity designs

Similar to natural experiments, discontinuity designs exploit sources of quasi-randomization originating in social and political processes. In contrast to the former, they rely on sharp jumps, or "discontinuities," in a continuous variable. The cutoff point determines whether a unit is exposed to the treatment or not, the idea being that treatment assignment is "as if at random" for units on either side of it. Elections are a typical example of such discontinuities because it is quite reasonable to assume that, in narrow elections, the outcome is due in large part to chance. While candidates that win by a landslide are likely very different from those who receive only a handful of vote, candidates on either side of the election threshold are probably similar in many respects.

Using these ideas, Eggers and Hainmueller (2009) compared the wealth at death of narrow winners and losers in British national elections and found that successful conservative candidates died with about 546,000 GBP, compared to about 298,000 GBP for candidates from the same party who were not elected. By contrast, the difference was much smaller for labour candidates, suggesting that the material benefits of serving in parliament differ across political parties. Gerber and Hopkins (2011) also relied on the random component of elections, but to examine the effects of partisanship on public policy at the local level. The comparison of 134 elections in 59 large American cities revealed that in most policy areas, changes in public spending were very similar regardless of whether a Republican or a Democrat narrowly won. The one exception were policing expenditures, which were higher under successful Republican candidates. These findings suggest that partisan effects are likely small at the local level.

Lalive and Zweimüller (2009) exploited a different type of discontinuity, namely, the date at which a longer parental leave entered into force in Austria, to estimate the effects of this policy on mothers' further childbearing and careers. Mothers giving birth after June 30, 1990 could benefit from a paid leave of two years, instead of one year under the policy in force until that date. Because of this sharp cutoff, the duration of the parental leave can be considered to be randomly assigned to mothers giving birth shortly before or after June 30. Indeed, the two groups were undistinguishable on many observed socio-economic characteristics such as age and work history and profile. The comparison of the two groups showed that a longer parental leaves causes women to have more additional children. It also reduces their employment and earnings, but only in the short term.

Sharp cutoffs like those found in elections and other settings generally offer quite convincing sources of quasi-randomization, even though researchers should carefully check whether actors could not be aware of the discontinuity and exploit it, like in the case of income tax thresholds (Green et al., 2009, 401). However, it is important to note that the causal effects estimated with this method apply only at the threshold and cannot be extrapolated to all units. Because, usually, only relatively few observations are sufficiently close to the threshold, the results produced by regression discontinuity designs apply to a specific subsample, which limit their external validity. Moreover, there are tradeoffs, but no clear guidelines, regarding the width of the window around the threshold Green et al. (2009). A larger window (and, thus, more observations) makes estimates more precise, but potentially biased by unobserved factors, while a smaller window reduces the bias but reduces the number of observation and, thus, the precision of the estimates.

4.3.3 Instrumental variables

Instrumental variables are factors that can be used to replace treatment variables for which the "as if at random" assumption does not hold (Sovey and Green, 2010). They have to meet three crucial assumptions. The first is relatively innocuous and states that the instrument and the treatment are correlated, after controlling for relevant covariates. The second and third are usually much more problematic. "Exclusion restriction" means that the instrument affects outcomes exclusively through its correlation with the treatment, that is, that it has no direct effect on the outcomes, while the "ignorability assumption" requires that the instrument is "as if at random." Thus, good instruments are those produced by some sort of quasi-experiment. Concretely, the estimation proceeds in two stages. In the first, the treatment variable is regressed on the instrument and the results are used to compute expected values for the treatment. In the second stage, these values replace the treatment in the main regression.

In a famous study, Acemoglu, Johnson and Robinson (2001) addressed the effects of institutions on economic development. A simple regression of development on institutions is likely to be inappropriate (even with many control variables) for two reasons. First, the causal relationship can arguably go both ways: better institutions cause higher economic development, but higher economic development can also cause better institutions. Second, similar to the example of women's quotas discussed earlier in this chapter, it is likely that countries with different degrees of economic development are different on many other dimensions as well. To circumvent these problems, Acemoglu, Johnson and Robinson (2001) employed mortality rates of European settlers (proxied by those of soldiers, bishops, and sailors) as an instrument for current institutions. The argument is that European powers set up different types of institutions depending on their ability to settle. If a region is hospitable, then European-style institutions with an emphasis on property rights and checks against government power are constructed while if it is not, colonizers set up "extractive states" with the purpose of transferring as much resources as possible from the colony. The analysis shows a strong association between current institutions, instrumented by settler mortality, and economic development, which corroborates the argument that a causal relationship is at play, rather than a mere correlation. An important caveat is the plausibility of the exclusion restriction, that is, the possibility that the effect of settler mortality on economic development could work through something else than institutions. For instance, the mortality rates of colonizers could be related to current diseases, which may have an impact on development. In this case, institutions would not be part of the causal chain. However, the authors argue convincingly that the causes of European deaths in the colonies (mainly malaria and yellow fever) are not likely to be connected with economic development because the indigenous populations have developed immunities against these diseases.

In another application, election day rainfall was used as instrument for turnout to estimate its effects on electoral outcomes in the United States (Hansford and Gomez, 2010). In effect, many studies have suggested that higher turnout is beneficial to leftist parties (or Democrats in the United States), but the problem is that many factors are likely to influence both the decision to vote and the vote itself at the same time. By contrast, the weather on election day likely affects the choice to go to the polling booth, but not the preference expressed in the vote.[3] Moreover, rainfall on a specific day can probably be considered an "as if at random" event. The analysis could confirm that, indeed, higher turnout causes a higher vote share for Democratic candidates.

Finally, in a study discussed already in Chapter 3, Kern and Hainmueller (2009) studied the effects of West German television on public support for the East Germany communist regime using a survey of East German teenagers. The survey included information for both the dependent (regime support) and treatment (exposure to West German television) variables. Because it is highly likely that people who watch a lot of West German programs have different predispositions towards the communist regime in the first place, the treatment cannot be considered "as if at random." However, while West German television reception was generally possible in East Germany, it was blocked in some regions (especially near Dresden) because of their topography. As long as living in Dresden per se was not directly related to regime support and that

[3] But see the Italian expression "Piove, governo ladro."

region was generally comparable with the rest of the country, living in Dresden can be used as an instrument for television exposure. The analysis showed that, quite counterintuitively, West German television caused greater support for the East German regime, possibly because East German citizens consumed it primarily for entertainment and not as a source of information.

Like the other approaches, instrumental variables come with their own set of problems (Sovey and Green, 2010). In fact, the list of potential issues is even longer because, in addition to the need to find a suitable "quasi-experiment," the instrument must fit within the model that is used in the estimation in a very specific way. Also, the results must be interpreted carefully because the causal effects estimates apply to a particular subset of units and are known as "local average treatment effects." In sum, if the right conditions are fulfilled instrumental variables are a valuable tool, but in practice their application is quite tricky.

5 Lessons for research design

If we take the statistical approach to causal inference seriously, the consequences for research design are wide ranging. The main lesson is that the design is the most important part of the research because it is at this stage that the possibility to credibly identify causal effects can be influenced. In fact, in the ideal-typical case of a "perfect" research design, that is, an experiment that is designed and implemented flawlessly, the analysis stage becomes almost trivial because it suffices to compare mean outcomes in the treatment and control groups. The sophistication of the methods used in the analysis must increase with imperfections in the research design in order to correct them ex post.

To illustrate, consider again the example of women's quotas and female representation in parliament (Tripp and Kang, 2008). The research design adopted by the authors, which is typical of cross-national quantitative studies, was simply to collect data on as many countries as possible for the dependent variable (percent of women in parliament), treatment variable (quotas), and control variables (countries' background characteristics). Here ends the design stage and begins the analysis, which, to produce credible causal estimates, needs to fix the basic problem that countries with and without quotas are not really comparable. As discussed in Section 4.1, standard regression tools and newer matching methods can help, but only up to a point. The fundamental problem is that they can adjust for the factors that we do observe, but not for those that we do not, which are virtually always an issue. Thus, ex post fixes are bound to be imperfect.

By contrast, the statistical approach to causal inference aims to fix things ex ante by constructing or finding suitable treatment and control groups in advance of the analysis. As we have seen, this goal can be achieved with different means. First, we can design our own experiments in the lab, in the field, or survey-based. That is, the treatment can be randomized by the researcher in an artificial setting, in the real world, or via the questions asked in a survey. Second, we can try to find constellations in which randomization is approximated without the direct intervention of the researcher. Natural experiments, discontinuity designs, and suitable instrumental variables are three options. In all these cases, most traction for causal inferences is gained through the way the comparison between treatment and control groups is configurated, not through the specific techniques used to analyze the data. The key benefit is that, if randomization is implemented properly or is approximated sufficiently in a real-world setting, it produces groups that are comparable not only for their observed, but also for their unobserved characteristics. This is a major advantage for the validity of causal inferences.

Thus, the quality of the research design is of the essence. The exacting requirement of a plausible "as if at random" assumption implies that downloading pre-packaged datasets and letting the computer do the counting is not enough, no matter how sophisticated the techniques. More creative solutions are required, and few will involve broad cross-national comparisons, for the simple reason that broad international comparisons will likely be, well, incomparable. In fact, none of the examples discussed in this chapter compared countries. Instead, they focused on specific within-country variations and used original data, often assembled with great efforts. Unfortunately, there are no clear guidelines for identifying promising comparisons. The criteria that the research design must meet are clear, but uncovering the right configuration in practice is an art more than a science.

Like all approaches, statistical research designs for causal inference must face trade offs. The most important is that between validity and relevance. A common criticism of this approach is that it leads to focus on small, tractable questions at the expense of big problems that are harder to study. It is undeniable that research in this tradition prioritizes internal over external validity. At the same time, the former is arguably a prerequisite for the second. In other

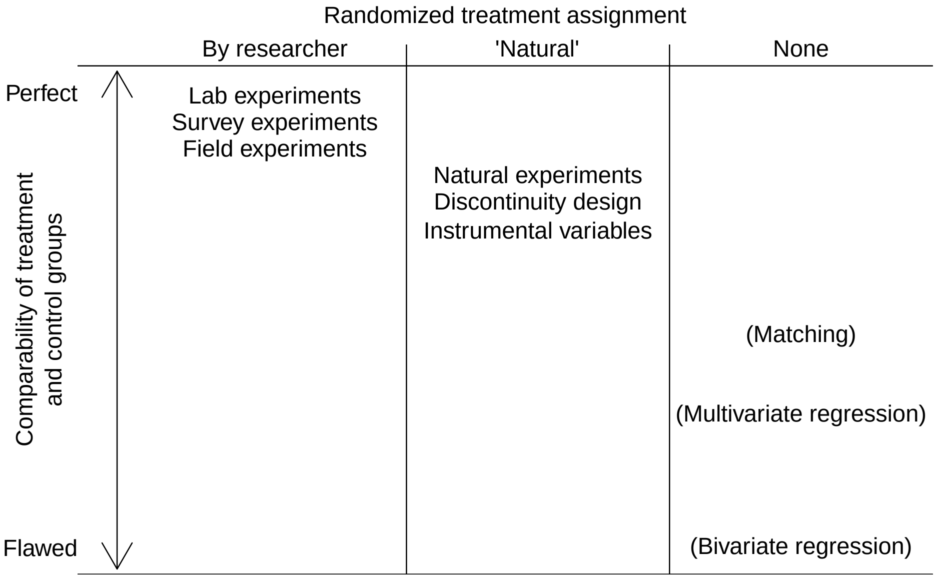

Figure 2: *A classification of statistical research designs for causal inference. Matching and regression are in parentheses because, strictly speaking, they are estimation techniques and not research designs.*

words, it does not make much sense to generalize findings that are not credible. Moreover, as Angrist and Pischke (2010) argue, external validity, or generalization, remains an important goal that can be achieved through the cumulation of well-designed but necessarily narrow studies. Finally, the examples discussed in this chapter studied problems such as the political salience of ethnicity, attitudes towards immigration, the consequences of direct democracy in comparison with representation, and the foreign influences of support for autocratic rule. These are all "big" questions and, even though each study individually did not provide definitive answers, they did supply convincing evidence on the causal effects in a specific setting. Other studies should try to replicate them in other contexts. If they are successful, then the external validity and generalizability of the findings will be strengthened.

6 Conclusion

Figure 2 summarizes the main points of this chapter. We can classify statistical research designs for causal inference along two dimensions. First, is the treatment assigned randomly, and if yes, how? Second, to what extent are the treated and control units comparable?

In the standard regression approach, supplemented or not by matching, there is no randomization and, typically, self-selection into the treatment. For instance, the same variables that explain why countries adopt women's quotas (the treatment) likely influence female representation in parliament (the outcome). The problem is bigger if these variables are not included in the analysis (bivariate regression) than if they are (multivariate regression), and matching can mitigate the problem further. However, there is no way around the fact that the adjustment can be made only for those variables that can be observed, but not for those that are unobserved. Therefore, the comparability of the treatment and control groups (countries with and without quotas) and, consequently, the validity of causal inferences, will be relatively limited.

By contrast, in experiments the treatment is randomized by researchers themselves and, in principle, the treated and control units will be highly comparable. Experiments can take place in the lab, in the field, and within surveys. Quasi-experiments can credibly make the assumption that the treatment is assigned "as if at random" because of a particular process occurring in the real world, without the researcher's intervention. The comparability of the treatment and control groups will in principle be quite high, significantly better than in the standard regression approach, but somewhat worse than in experiments. The validity of the causal inferences will vary accordingly.

In this context, an important tradeoff is that between complexity or realism of the research question and reliability of the causal estimated. To achieve the latter, statistical research designs narrow down complex theoretical and/or empirical questions to smaller, tractable questions. These research designs can produce valid estimates of causal relationships, but many different analyses are necessary to give the full picture of a complex phenomenon. By contrast, other research designs discussed in this book put the emphasis on a holistic view of causal processes, but at the cost of validity.

To conclude, the statistical approach emphasizes the importance of research design for valid causal inferences. The primary concern is the construction of comparable treatment and control groups. This will be difficult with standard cross-national datasets. Instead, researchers should produce their own experiments or look for configurations in the real world that can approximate them.

7 Questions

1. Read closely five articles making causal arguments in your field of study. To what extent do they correspond to a "causes-of-effects" or "effects-of-causes" perspective?

2. For each of the five articles, reframe the causal claims using the potential outcomes framework and construct the equivalent of Table 1.

3. Read five articles making causal arguments using standard regression methods. To what extent can the findings actually be interpreted causally?

4. Think of a specific research question. What would be the ideal experiment to test the causal argument? Now try to develop a research design that can approximate it as much as possible in practice.

5. Read closely five of the articles cited as examples in this chapter (or other articles of your choice) and assess them with respect to the trade off between the validity of the causal inference and the relevance or importance of the findings.

References

Acemoglu, Daron, Simon Johnson and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91(5):1369–1401.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist Companion*. Princeton: Princeton University Press.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2):3–30.

Barabas, Jason and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104(02):226–242.

Bhavnani, Rikhil R. 2009. "Do Electoral Quotas Work After They Are Withdrawn? Evidence from a Natural Experiment in India." *American Political Science Review* 103(1):23–35.

Correll, Shelley J. 2004. "Constraints into Preferences: Gender, Status, and Emerging Career Patterns." *American Sociological Review* 69:93–113.

Druckman, James N, Donald P Green, James H. Kuklinski and Arthur Lupia. 2006. "The growth and development of experimental research in political science." *American Political Science Review* 100(4):627–635.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2):282–293.

Dunning, Thad and Lauren Harrison. 2010. "Cross-cutting Cleavages and Ethnic Voting: An Experimental Study of Cousinage in Mali." *American Political Science Review* 104(1):21.

Eggers, Andrew and Jens Hainmueller. 2009. "MPs for Sale? Returns to Office in Postwar British Politics." *American Political Science Review* 103(4):513–533.

Gaines, Brian J., James H. Kuklinski and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(1):1–20.

Gelman, Andrew. 2011. "Causality and Statistical Learning." *American Journal of Sociology* 117(3):955–966.

Gerber, Alan S. and Donald P. Green. 2008. Field Experiments and Natural Experiments. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier. Oxford: Oxford University Press pp. 357–381.

Gerber, Alan S., James G. Gimpel, Donald P Green and Daron R. Shaw. 2011. "How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105(1):135–150.

Gerber, Elisabeth R and Daniel J Hopkins. 2011. "When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy." *American Journal of Political Science* 55(2):326–339.

Green, Donald P., Terence Y. Leong, Holger Lutz Kern, Alan S. Gerber and Christopher W. Larimer. 2009. "Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks." *Political Analysis* 17(4):400–417.

Hainmueller, Jens and Michael J. Hiscox. 2010. "Attitudes toward Highly Skilled and Low-skilled Immigration: Evidence from a Survey Experiment." *American Political Science Review* 104(1):61.

Hansford, Thomas G. and Brad T. Gomez. 2010. "Estimating the Electoral Effects of Voter Turnout." *American Political Science Review* 104(2):268–288.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Humphreys, Macartan and Jeremy M Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12(1):367–378.

Hyde, Susan D. 2007. "The observer effect in international politics: Evidence from a natural experiment." *World Politics* 60:37–63.

Kern, Holger Lutz and Jens Hainmueller. 2009. "Opium for the Masses: How Foreign Free Media Can Stabilize Authoritarian Regimes." Cornell University and Harvard University.

Lalive, Rafael and Josef Zweimüller. 2009. "How Does Parental Leave Affect Fertility and Return to Work? Evidence from Two Natural Experiments." *Quarterly Journal of Economics* 124(3):1363–1402.

Linos, Katerina. 2011. "Diffusion through Democracy." *American Journal of Political Science* 55(3):678–695.

Mahoney, James and Gary Goertz. 2006. "A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research." *Political Analysis* 14:227–249.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference. Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

Morton, Rebecca B. and Kenneth C. Williams. 2008. Experimentation in Political Science. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier. Oxford: Oxford University Press pp. 339–356.

Morton, Rebecca B. and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab.* Cambridge: Cambridge University Press.

Olken, Benjamin A. 2010. "Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia." *American Political Science Review* 104(2):243–267.

Ragin, Charles. 1987. *The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies.* Berkeley: University of California Press.

Sovey, Allison J. and Donald P. Green. 2010. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

Tripp, Aili Mari and Alice Kang. 2008. "The Global Impact of Quotas: On the Fast Track to Increased Female Legislative Representation." *Comparative Political Studies* 41(3):338–361.