Cambridge Handbook of Experimental Political Science

Edited by James N. Druckman, Donald P. Green, James H. Kuklinski, Arthur Lupia

Book DOI: http://dx.doi.org/10.1017/CBO9780511921452

Online ISBN: 9780511921452

Hardback ISBN: 9780521192125

Paperback ISBN: 9780521174558

Chapter

2 - Experiments pp. 15-26

Chapter DOI: http://dx.doi.org/10.1017/CBO9780511921452.002

Cambridge University Press

CHAPTER 2

Experiments

An Introduction to Core Concepts

James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia

The experimental study of politics has grown explosively in the past two decades. Part of that explosion takes the form of a dramatic increase in the number of published articles that use experiments. Perhaps less evident, and arguably more important, experimentalists are exploring topics that would have been unimaginable only a few years ago. Laboratory researchers have studied topics ranging from the effects of media exposure (Iyengar and Kinder 1987) to the conditions under which groups solve collective action problems (Ostrom, Walker, and Gardner 1992), and, at times, have identified empirical anomalies that produced new theoretical insights (McKelvey and Palfrey 1992). Some survey experimenters have developed experimental techniques to measure prejudice (Kuklinski, Cobb, and Gilens 1997) and its effects on support for policies such as welfare or affirmative action (Sniderman and Piazza 1995); others have explored the ways in which framing, information, and decision cues influence voters' policy preferences and support for public officials (Druckman 2004; Tomz 2007). And although the initial wave of field experiments focused on the effects of campaign communications on turnout and voters' preferences (Eldersveld 1956; Gerber and Green 2000; Wantchekon 2003), researchers increasingly use field experiments and natural experiments to study phenomena as varied as election fraud (Hyde 2009), representation (Butler and Nickerson 2009), counterinsurgency (Lyall 2009), and interpersonal communication (Nickerson 2008).

With the rapid growth and development of experimental methods in political science come a set of terms and concepts that political scientists must know and understand. In this chapter, we review concepts and definitions that often appear in the *Handbook* chapters. We also highlight features of experiments that are unique to political science.

We thank Holger Kern for helpful comments.

1. What Is an Experiment?

In contrast to modes of research that address descriptive or interpretive questions,

researchers design experiments to address causal questions. A causal question invites a comparison between two states of the world: one in which some sort of intervention is administered (a treated state, i.e., exposing a subject to a stimulus) and another in which it is not (an untreated state). The *fundamental problem of causal inference* arises because we cannot simultaneously observe a person or entity in its treated and untreated states (Holland 1986).

Consider, for example, the causal effect of viewing a presidential debate. Rarely are the elections of 1960, 1980, 1984, or 2000 recounted without mentioning the critical role that debates played in shaping voter opinion. What is the basis for thinking that viewing a presidential debate influences the public's support for the candidates? We do not observe how viewers of the debate would have voted had they not seen the debate. We do not observe how nonviewers would have voted had they watched (Druckman 2003). Nature does not provide us with the observations we would need to make the precise causal comparisons that we seek.

Social scientists have pursued two empirical strategies to overcome this conundrum: observational research and experimental research. *Observational research* involves a comparison between people or entities subjected to different treatments (at least, in part, of their own choosing). Suppose that some people watched a presidential debate, whereas others did not. To what extent can we determine the effect of debate watching by comparing the postdebate behaviors of viewers and nonviewers? The answer depends on the extent to which viewers and nonviewers are otherwise similar. It might be that most debate watchers already supported one candidate, whereas most nonwatchers favored the other. In such cases, observed differences between the postdebate opinions of watchers and nonwatchers could stem largely from differences in the opinions they held before the debate even started. Hence, to observe that viewers and nonviewers express different views about a candidate after a debate does not say unequivocally that watching the debate caused these differences.

In an effort to address such concerns, observational researchers often attempt to compare treated and untreated people only when they share certain attributes, such as age or ideology. Researchers implement this general approach in many ways (e.g., multiple regression analysis, case-based matching, case control methodology), but all employ a similar underlying logic: find a group of seemingly comparable observations that have received different treatments, then base the causal evaluation primarily or exclusively on these observations.

Such approaches often fail to eliminate comparability problems. There might be no way to know whether individuals who look similar in terms of a (usually limited) set of observed attributes would in fact have responded identically to a particular treatment. Two groups of individuals who look the same to researchers could differ in unmeasured ways (e.g., openness to persuasion). This problem is particularly acute when people self-select into or out of a treatment. Whether people decide to watch or not watch a debate, for example, might depend on unmeasured attributes that predict which candidate they support (e.g., people who favor the front-running candidate before the debate might be more likely to watch the debate than those who expect their candidate to lose).

Experimental research differs from observational research in that the entities under study are randomly assigned to different treatments. Here, *treatments* refer to potentially causal interventions. For example, an experimenter might assign some people to watch a debate (one treatment) and assign others to watch a completely different program (a second treatment). Depending on the experimental design, there may be a *control group* that does not receive a treatment (e.g., they are neither told to watch nor discouraged from watching the debate) and/or an alternative treatment group (e.g., they are told to watch a different show or a different part of the debate). *Random assignment*

means that each entity being studied has an equal chance to be in a particular treatment condition.[1]

Albertson and Lawrence (2009) and Mullainathan, Washington, and Azari (2010), for example, discuss experiments with *encouragement designs* in which the researcher randomly encourages some survey respondents to view an upcoming candidate debate (treatment group) and neither encourages or discourages others (control group). After the debate, the researcher conducts a second interview with both groups in order to ascertain whether they watched the debate and to measure their candidate preferences.

How does random assignment overcome the fundamental problem of causal inference? Suppose for the time being that everyone who was encouraged to view the debate did so and that no one watched unless encouraged. Although we cannot observe a given individual in both his or her treated and untreated states, random assignment enables the researcher to estimate the *average treatment effect*. Prior to the intervention, the randomly assigned treatment and control groups have the same expected responses to viewing the debate. Apart from random sampling variability, in other words, random assignment provides a basis for assuming that the control group behaves as the treatment group would have behaved had it not received the treatment (and vice versa). By comparing the average outcome in the treatment group to the average outcome in the control group, the experimental researcher estimates the average treatment effect. Moreover, the researcher can perform statistical tests to clarify whether the differences between groups occurred simply by chance (sampling variability) rather than as a result of experimental treatments.

When we speak of an experiment in this *Handbook*, we mean a study in which the units of observation (typically, subjects, or human participants in an experiment) are randomly assigned to different treatment or control groups (although see note 2). Experimental studies can take many forms. It is customary to classify randomized studies according to the settings in which they take place: a *lab experiment* involves an intervention in a setting created and controlled by the researcher; a *field experiment* takes place in a naturally occurring setting; and a *survey experiment* involves an intervention in the course of an opinion survey (which might be conducted in person, over the phone, or via the web). This classification scheme is not entirely adequate, however, because studies often blend different aspects of lab, field, and survey experiments. For example, some experiments take place in lab-like settings, such as a classroom, but require the completion of a survey that contains the experimental treatments (e.g., the treatments might entail providing individuals with different types of information about an issue).

2. Random Assignment or Random Sampling?

When evaluating whether a study qualifies as an experiment, by our definition, random *assignment* should not be confused with random *sampling*. Random sampling refers to a procedure by which participants are selected for certain kinds of studies. A common random sampling goal is to choose participants from a broader population in a way that gives every potential participant the same probability of being selected into the study. Random assignment differs. It does not require that participants be drawn randomly from some larger population. Experimental participants might come from undergraduate courses or from particular towns. The key requirement is that a random procedure, such as a coin flip, determines whether they receive a particular treatment. Just as an experiment does not require a random sample, a study of a random sample need not be an

1 In the social sciences, in contrast to the physical sciences, experiments tend to involve use of random assignment to treatment conditions. Randomly assigned treatments are one type of "independent variable." Another type comprises "covariates" that are not randomly assigned but nonetheless predict the outcome.

experiment. A survey that merely asks a random sample of adults whether they watched a presidential debate might be a fine study, but it is not an experimental study of the effects of debate viewing because watching or not watching the debate was not randomly assigned.

The typical social science experiment uses a *between-subjects design*, insofar as the researcher randomly assigns participants to distinct treatment groups. An alternative approach is the *within-subjects design* in which a given participant is observed before and after receiving a treatment (e.g., there is no random assignment between subjects). Intuitively, the within-subjects design seems to overcome the fundamental problem of causal inference; in practice, it is often vulnerable to confounds – meaning unintended and uncontrolled factors that influence the results. For example, suppose that a researcher measures subjects' attitudes toward a candidate before they watch a debate, and then again after they have watched it, to determine whether the debate changed their attitudes. If subjects should hear attitude-changing news about the candidate after the first measurement and prior to the second, or if simply filling out the predebate questionnaire induces them to watch the debate differently than they otherwise would have watched, a comparison of pre- and postattitudes will produce misleading conclusions about the effect of the debate.[2]

3. Internal and External Validity

Random assignment enables the researcher to formulate the appropriate comparisons, but random assignment alone does not ensure that the comparison will speak convincingly to the original causal question. The theoretical interpretation of an experimental result is a matter of *internal validity* – "did in fact the experimental stimulus [e.g., the debate] make some significant difference [e.g., in attitude toward the candidate] in this specific instance" (Campbell 1957, 297).[3] In the preceding example, the researcher seeks to gauge the causal effect of viewing a televised debate; however, if viewers of the debate are inadvertently exposed to attitude-changing news as well, then the estimated effect of viewing the debate will be conflated with the effect of hearing the news.

The interpretation of the estimated causal effect also depends on what the control group receives as a treatment. If, in the previous example, the control group watches another television program that airs campaign commercials, the researcher must understand the treatment effect as the relative influence of viewing debates compared to viewing commercials.[4] This comparison differs from a comparison of those who watch a debate with those who, experimentally, watch nothing.

More generally, every experimental treatment entails subtle nuances that the researcher must know, understand, and explicate. Hence, in the preceding example, he or she must judge whether the causative agent was viewing a debate per se, any 90-minute political program, or any political program of any length. Researchers can, and should, conduct

2 Natural scientists frequently use within-subjects designs because they seldom contend with problems of memory and anticipation when working with "subjects" such as electrons. Clearly, natural scientists conduct "experiments" (with interventions) even if they do not employ between-subjects random assignment. Social scientists, confronted as they are by the additional complexities of working with humans, typically rely on between-subjects experimental designs, where randomization ensures that the experimental groups are, in expectation, identical.

Randomization is unnecessary when subjects are effectively identical. In economics (and hence some of the work discussed in this *Handbook*), participants sometimes are not randomly assigned on the assumption that they all respond the same way to the incentives provided by the experimenter (Guala 2005, 79; Morton and Williams 2010, 28–29).

3 Related to internal validity is statistical conclusion validity, defined as "the validity of inferences about the correlation (covariation) between treatment and outcome" (Shadish et al. 2002, 38). Statistical conclusion validity refers specifically and solely to the "appropriate use of statistics to infer whether the presumed independent and dependent variables covary," and not at all to whether a true causal relationship exists (37).

4 Internal validity is a frequent challenge for experimental research. For this reason, experimental scholars often administer *manipulation checks*, evaluations that document whether subjects experience the treatment as intended by the experimenter.

multiple experiments or experiments with a wide array of different conditions in an effort to isolate the precise causative agent; however, at the end of the day, they must rely on theoretical stipulations to decide which idiosyncratic aspects of the treatment are relevant and explain why they, and not others, are relevant.

Two aspects of experimental implementation that bear directly on internal validity are *noncompliance* and *attrition*. Noncompliance occurs when those assigned to the treatment group do not receive the treatment, or when those assigned to the control group inadvertently receive the treatment (e.g., those encouraged to watch do not watch or those not encouraged do watch). In this case, the randomly assigned groups remain comparable, but the difference in their average outcomes measures the effect of the experimental assignment rather than actually receiving the treatment. The Appendix to this chapter describes how to draw causal inferences in such circumstances.

Attrition involves the failure to measure outcomes for certain subjects (e.g., some do not report their vote preference in the follow-up) and is particularly problematic when it afflicts some experimental groups more than others. The danger is that attrition reveals something about the potential outcomes of those who drop out of the study. For example, if debate viewers become more willing than nonviewers to participate in a postdebate interview and if viewing the debate changes subjects' candidate evaluations, comparisons between treatment and control group could be biased. Sometimes researchers unwittingly contribute to the problem of differential attrition by exerting more effort to gather outcome data from one of the experimental groups or by expelling participants from the study if they fail to follow directions when receiving the treatment.

A related concern for experimental researchers is *external validity*. Researchers typically conduct experiments with an eye toward questions that are bigger than "What is the causal effect of the treatment on this particular group of people?" For example, they may want to provide insight about voters generally, despite having data on relatively few voters. How far one can generalize from the results of a particular experiment is a question of *external validity*: the extent to which the "causal relationship holds over variations in persons, settings, treatments, and outcomes" (Shadish, Cook, and Campbell 2002, 83).[5]

As suggested in the Shadish et al. quote, external validity covers at least four aspects of experimental design: whether the participants resemble the actors who are ordinarily confronted with these stimuli, whether the context (including the time) within which actors operate resembles the context (and time) of interest, whether the stimulus used in the study resembles the stimulus of interest in the world, and whether the outcome measures resemble the actual outcomes of theoretical or practical interest. The fact that several criteria come into play means that experiments are difficult to grade in terms of external validity, particularly because the external validity of a given study depends on what kinds of generalizations one seeks to make.

Consider the external validity of our example of the debate-watching encouragement experiment. The subjects in encouragement studies come from random samples of the populations of adults or registered voters. Random sampling bolsters the external validity of the study insofar as the people in the survey better reflect the target population. However, if certain types of people comply with encouragement instructions more than others, then our post-treatment inferences will depend on whether the average effect among those who comply with the treatment resembles the average effect among those groups to which we hope to generalize.

A related concern in such experiments is whether the context and time at which participants watch the debate resembles settings to which the researcher hopes to generalize. Are the viewers allowed to ignore the debate and read a magazine if they want (as they could outside the study)? Are they watching with the same types of people they would

5 Related is construct validity, which is "the validity of inferences about the higher order constructs that represent sampling particulars" (Shadish et al. 2002, 38).

watch with outside the study? There also are questions about the particular debate program used in the study (e.g., the stimulus): does it typify debates in general? To the extent that it does not, it will be more difficult to make claims about debate viewing that are regarded as externally valid. Before generalizing from the results of such an experiment, we would need to know more about the tone, content, and context of the debate.[6]

Finally, suppose our main interest is in how debate viewing affects Election Day behaviors. If we want to understand how exposure to debates influences voting, then a questionnaire given on Election Day might be a better measurement than one taken immediately after the debate and well before the election; that is, behavioral intentions may change after the debate but before the election.

Whether any of these concerns make a material difference to the external validity of an experimental finding can be addressed as part of an extended research program in which scholars vary relevant attributes of the research design, such as the subjects targeted for participation, the alternative viewing (or reading) choices available (to address the generalizability of effects from watching a particular debate in a certain circumstance), the types of debates watched, and the timing of postdebate interviews. A series of such experiments could address external validity concerns by gradually assessing how treatment effects vary depending on different attributes of experimental design.

4. Documenting and Reporting Relationships

When researchers detect a statistical relationship between a randomly assigned treatment and an outcome variable, they often want to probe further to understand the mechanisms by which the effect is transmitted. For example, having found that watching a televised debate increased the likelihood of voting, they ask why it has this effect. Is it because viewers become more interested in the race? Do they feel more confident about their ability to cast an intelligent vote? Do debates elevate their feelings of civic duty? Viewing a debate could change any of these *mediating variables*.

Assessing the extent to which potential mediating variables explain an experimental effect can be challenging. Analytically, a single random assignment (viewing a debate vs. not viewing) makes it difficult, if not impossible, to isolate the mediating pathways of numerous intervening variables. To clarify such effects, a researcher needs to design several experiments, all with different kinds of treatments. In the debate example, a researcher could ask subjects to watch different kinds of debates, with some treatments likely to affect interest in the race and others to heighten feelings of civic duty. Indeed, an extensive series of experiments might be required before a researcher can make convincing causal claims about causal pathways.

In addition to identifying mediating variables, researchers often want to understand the conditions under which an experimental treatment affects an important outcome. For example, do debates only affect (or affect to a greater extent) political independents? Do debates matter only when held in close proximity to Election Day? These are questions about *moderation*, wherein the treatment's effect on the outcome differs across levels of other variables (e.g., partisanship, timing of debate [see Baron and Kenny 1986]). Documenting moderating relationships typically entails the use of statistical interactions between the moderating variable and the treatment. This approach, however, requires sufficient variance on the moderating variable. For example, to evaluate whether debates affect only independents, the subject population must include sufficient numbers of otherwise comparable independents and nonindependents.

6 This is related to the aforementioned internal validity concern about whether the content of the debate itself caused the reaction, or whether any such programming would have caused it. The internal validity concern is about the causal impact of the presumed stimulus – is the cause what we believe it is (e.g., the debate and not any political programming)? The external validity concern is about whether that causal agent reflects the set of causal variables to which we hope to infer (e.g., is the content of the debate representative of presidential debates?).

In practice, pinpointing mediators and moderators often requires theoretical guidance and the use of multiple experiments representing distinct conditions. This gets at one of the great advantages of experiments – they can be *replicated* and extended in order to form a body of related studies. Moreover, as experimental literatures develop, they lend themselves to *meta-analysis*, a form of statistical analysis that assesses the conditions under which effects are large or small (Borenstein et al. 2009). Meta-analyses aggregate experiments on a given topic into a single dataset and test whether effect sizes vary with certain changes in the treatments, subjects, context, or manner in which the experiments were implemented. Meta-analysis can reveal statistically significant treatment effects from a set of studies that, analyzed separately, would each generate estimated treatment effects indistinguishable from zero. Indeed, it is this feature of meta-analysis that argues against the usual notion that one should always avoid conducting experiments with low *statistical power*, or a low probability of rejecting the null hypothesis of no effect (when there is in fact an effect).[7] A set of low power studies taken together might have considerable power, but if no one ever launches a low power study, this needed evidence cannot accumulate (for examples of meta-analyses in political science, see Druckman 1994; Lau et al. 1999).[8]

Publication bias threatens the accumulation of experimental evidence through meta-analysis. Some experiments find their way into print more readily than others. Those that generate statistically significant results and show that the effect of administering a treatment is clearly nonzero are more likely to be deemed worthy of publication by journal reviewers, editors, and even authors themselves. If statistically significant positive results are published and weaker results are not, then the published literature will give a distorted impression of a treatment's influence. A meta-analysis of results that have been published selectively might be quite misleading. For example, if only experiments documenting that debates affect voter opinion survive the publication process and those that report no effects are never published, then the published literature may provide a skewed view of debate effects. For this reason, researchers who employ meta-analysis should look for symptoms of publication bias, such as the tendency for smaller studies to generate larger treatment effects.

As the discussions of validity and publication bias suggest, experimentation is no panacea.[9] The interpretation of experimental results requires intimate knowledge of how and under what conditions an experiment was conducted and reported. For this reason, it is incumbent on experimental researchers to give a detailed account of the key features of their studies, including 1) who the subjects are and how they came to participate in the study; 2) how the subjects were randomly assigned to experimental groups; 3) what treatments each group received; 4) the context in which each group received treatments; 5) the outcome measures; and 6) all procedures used to preserve comparability between treatment and control groups, such as outcome measurement that is blind to participants' experimental assignments and the management of noncompliance and attrition.

5. Ethics and Natural Experiments

Implementing experiments in ways that speak convincingly to causal questions is important

7 Statistical power refers to the probability that a researcher will reject the null hypothesis of no effect when the alternative hypothesis is indeed true.

8 Early lab and field studies of the mass media fall into this category. Iyengar, Peters, and Kinder's (1982) influential lab study of television news had less than twenty subjects in some of the experimental conditions. Panagopoulos and Green's (2008) study of radio advertising comprised a few dozen mayoral elections. Neither produced overwhelming statistical evidence on its own, but both have been bolstered by replications.

9 The volume does not include explicit chapters on meta-analysis or publication bias, reflecting, in part, the still relatively recent rise in experimental methods (i.e., in many areas, there is not yet a sufficient accumulation of evidence). We imagine these topics will soon receive considerably more attention within political science.

and challenging. Experiments that have great clarifying potential can also be expensive and difficult to orchestrate, particularly in situations where the random assignment of treatments means a sharp departure from what would ordinarily occur. For experiments on certain visible or conflictual topics, ethical problems might also arise. Subjects might either be denied a treatment that they would ordinarily seek or be exposed to a treatment that they would ordinarily avoid. Even if the ethical problems are manageable, such situations might also require researchers to garner potential subjects' explicit consent to participate in the experimental activities. Subjects might refuse to consent, or the consent form might prompt them to think or behave in ways they otherwise would not – in both instances, challenging the external validity of the experiment. Moreover, some studies include deception, an aspect of experimental design that raises not only ethical qualms, but also practical concerns about jeopardizing the credibility of the experimental instructions in future experiments.

Hence, the creative spark required of a great experimental study is not just how to test an engaging hypothesis, but how to conduct a test while effectively managing practical and ethical constraints. In some cases, researchers address such practical and ethical hurdles by searching for and taking advantage of random assignments that occur naturally in the world. These *natural experiments* include instances where random lotteries determine which men are drafted for military service (e.g., Angrist 1990), which incoming legislators enjoy the right to propose legislation (Loewen, Koop, and Fowler 2009), or which Pakistani Muslims obtain visas allowing them to make the pilgrimage to Mecca (Clingingsmith, Khwaja, and Kremer 2009). The term *natural experiment* is sometimes defined more expansively to include events that happen to some people and not others, but the happenstance is not random. The adequacy of this broader definition is debatable; however, when the mechanism determining whether people are exposed to a potentially relevant stimulus is sufficiently random, then these natural experiments can provide scholars with an opportunity to conduct research on topics that would ordinarily be beyond an experimenter's reach.

6. Conclusion

That social science experiments take many forms reflects different judgments about how best to balance various research aims. Some scholars prefer laboratory experiments to field experiments on the grounds that the lab offers the researcher tighter control over the treatment and how it is presented to subjects. Others take the opposite view on the grounds that generalization will be limited unless treatments are deployed, and outcomes assessed, unobtrusively in the field. Survey experiments are sometimes preferred on the grounds that a large and representative sample of people can be presented with a broad array of different stimuli in an environment where detailed outcome measures are easily gathered. Finally, some scholars turn to natural experiments in order to study historical interventions or interventions that could not, for practical or ethical reasons, be introduced by researchers.

The diversity of experimental approaches reflects in part different tastes about which research topics are most valuable, as well as ongoing debates within the experimental community about how best to attack particular problems of causal inference. Thus, it is difficult to make broad claims about "the right way" to run experiments in many substantive domains. In many respects, experimentation in political science is still in its infancy, and it remains to be seen which experimental designs, or combinations of designs, provide the most reliable political insights. That said, a good working knowledge of this chapter's basic concepts and definitions can further understanding of the reasons behind the dramatic growth in the number and scope of experiments in political science, as well as the ways in which others are likely to evaluate and learn from the experiments that a researcher develops.

Appendix: Introduction to the Neyman-Rubin Causal Model

The logic underlying randomized experiments is often explicated in terms of a notational system that has its origins in Neyman (1923) and Rubin (1974). For each individual i, let Y_0 be the outcome if i is not exposed to the treatment and Y_1 be the outcome if i is exposed to the treatment. The treatment effect is defined as

$$\tau_i = Y_{i1} - Y_{i0}. \tag{1}$$

In other words, the treatment effect is the difference between two potential states of the world: one in which the individual receives the treatment and another in which the individual does not. Extending this logic from a single individual to a set of individuals, we may define the average treatment effect (ATE) as follows:

$$ATE = E(\tau_i) = E(Y_{i1}) - E(Y_{i0}). \tag{2}$$

The concept of the average treatment effect implicitly acknowledges the fact that the treatment effect may vary across individuals. The value of τ_i may be especially large, for example, among those who seek out a given treatment. In such cases, the average treatment effect in the population may be quite different from the average treatment effect among those who actually receive the treatment.

Stated formally, the concept of the average treatment effect among the treated may be written as

$$ATT = E(\tau_i|T_i = 1) = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1), \tag{3}$$

where $T_i = 1$ when a person receives a treatment. To clarify the terminology, $Y_{i1}|T_i = 1$ is the outcome resulting from the treatment among those who are actually treated, whereas $Y_{i0}|T_i = 1$ is the outcome that would have been observed in the absence of treatment among those who are actually treated. By comparing Equations (2) and (3), we see that the average treatment effect need not be the same as the treatment effect among the treated.

This framework can be used to show the importance of random assignment. When treatments are randomly administered, the group that receives the treatment ($T_i = 1$) has the same expected outcome that the group that does not receive the treatment ($T_i = 0$) would if it were treated:

$$E(Y_{i1}|T_i = 1) = E(Y_{i1}|T_i = 0). \tag{4}$$

Similarly, the group that does not receive the treatment has the same expected outcome, if untreated, as the group that receives the treatment, if it were untreated:

$$E(Y_{i0}|T_i = 0) = E(Y_{i0}|T_i = 1). \tag{5}$$

Equations (4) and (5) follow from what Holland (1986) terms the independence assumption because the randomly assigned value of T_i conveys no information about the potential values of Y_i. Equations (2), (4), and (5) imply that the average treatment effect may be written as

$$ATE = E(\tau_i) = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0). \tag{6}$$

Because $E(Y_{i1}|T_i = 1)$ and $E(Y_{i0}|T_i = 0)$ may be estimated directly from the data, Equation (6) suggests a solution to the problem of causal inference. To estimate an average treatment effect, we simply calculate the difference between two sample means: the average outcome in the treatment group minus the average outcome in the control group. This estimate is unbiased in the sense that, on average across hypothetical replications of the same experiment, it reveals the true average treatment effect.

Random assignment further implies that independence will hold not only for Y_i, but also for any variable X_i that might be measured prior to the administration of the treatment. For example, subjects' demographic

attributes or their scores on a pre-test are presumably independent of randomly assigned treatment groups. Thus, one expects the average value of X_i in the treatment group to be the same as in the control group; indeed, the entire distribution of X_i is expected to be the same across experimental groups. This property is known as *covariate balance*. It is possible to gauge the degree of balance empirically by comparing the sample averages for the treatment and control groups.

The preceding discussion of causal effects skipped over two further assumptions that play a subtle but important role in experimental analysis. The first is the idea of an *exclusion restriction*. Embedded in Equation (1) is the idea that outcomes vary as a function of receiving the treatment per se. It is assumed that assignment to the treatment group only affects outcomes insofar as subjects receive the treatment. Part of the rationale for using blinded placebo groups in experimental design is the concern that subjects' knowledge of their experimental assignment might affect their outcomes. The same may be said for double-blind procedures: when those who implement experiments are unaware of subjects' experimental assignments, they cannot intentionally or inadvertently alter their measurement of the dependent variable.

A second assumption is known as the *stable unit treatment value assumption* (SUTVA). In the notation used previously, expectations such as $E(Y_{i1}|T_i = t_i)$ are all written as if the expected value of the treatment outcome variable Y_{i1} for unit i only depends on whether the unit gets the treatment (whether t_i equals one or zero). A more complete notation would allow for the consequences of treatments T_1 through T_n administered to other units. It is conceivable that experimental outcomes might depend on the values of $t_1, t_2, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n$ as well as the value of t_i:

$$E(Y_{i1}|T_1 = t_1,\ T_2 = t_2,\ \ldots,\ T_{i-1} = t_{i-1},\ T_i = t_i,\ T_{i+1} = t_{i+1}, \ldots,\ T_n = t_n).$$

By ignoring the assignments to all other units when we write this as $E(Y_{i1}|T_i = t_i)$, we assume away spillovers (or multiple forms of the treatment) from one experimental subject to another.

Noncompliance

Sometimes only a subset of those who are assigned to the treatment group is actually treated, or a portion of the control group receives the treatment. When those who get the treatment differ from those who are assigned to receive it, an experiment confronts a problem of *noncompliance*. In experimental studies of get-out-the-vote canvassing, for example, noncompliance occurs when some subjects who were assigned to the treatment group remain untreated because they are not reached (see Gerber et al. 2010).

How experimenters approach the problem of noncompliance depends on their objectives. Those who want to gauge the effectiveness of an outreach program may be content to estimate the *intent-to-treat effect*, that is, the effect of being randomly assigned to the treatment. The intent-to-treat effect is essentially a blend of two aspects of the experimental intervention: the rate at which the assigned treatment is actually delivered to subjects and the effect it has on those who receive it. Some experimenters are primarily interested in the latter. Their aim is to measure the effects of the treatment on *compliers*, people who receive the treatment if and only if they are assigned to the treatment group.

When there is noncompliance, a subject's group assignment, Z_i, is not equivalent to T_i. Define a subset of the population, called "compliers," who get the treatment if and only if they are assigned to the treatment. Compliers are subjects for whom $T_i = 1$ when $Z_i = 1$ and for whom $T_i = 0$ when $Z_i = 0$. Note that whether a subject is a complier is a function of both a subject's characteristics and the particular features of the experiment; it is not a fixed attribute of a subject.

When treatments are administered exactly according to plan ($Z_i = T_i,\ \forall_i$), the average causal effect of a randomly assigned treatment can be estimated simply by comparing mean

treatment group outcomes and mean control group outcomes. What can be learned about treatment effects when there is noncompliance? Angrist et al. (1996) present a set of sufficient conditions for estimating the average treatment effect among the subgroup of subjects who are compliers. Here we present a description of the assumptions for estimating the average treatment effect for the compliers.

To estimate the average treatment effect among compliers, we must assume that assignment Z is random. We must also make four additional assumptions: the exclusion restriction, SUTVA, monotonicity, and a nonzero causal effect of the random assignment. The exclusion restriction implies that the outcome for a subject is a function of the treatment they receive but is not otherwise influenced by their assignment to the treatment group. SUTVA implies that a subject's outcomes depend only on the subject's own treatment and treatment assignment and not on the treatments assigned or received by any other subjects. Monotonicity means that there are no *defiers*, that is, no subjects who would receive the treatment if assigned to the control group and who would not receive the treatment if assigned to the treatment group. The final assumption is that the random assignment has some effect on the probability of receiving the treatment. With these assumptions in place, the researcher may estimate the average treatment effect among compliers in a manner that will be increasingly accurate as the number of observations in the study increases. Thus, although the problem of experimental crossover constrains a researcher's ability to draw inferences about the average treatment effect among the entire population, accurate inferences can often be obtained with regard to the average treatment effect among compliers.

References

Albertson, Bethany, and Adria Lawrence. 2009. "After the Credits Roll: The Long-Term Effects of Educational Television on Public Knowledge and Attitudes." *American Politics Research* 37: 275–300.

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

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444–55.

Baron, Reuben M., and David A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51: 1173–82.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. London: Wiley.

Butler, Daniel M., and David W. Nickerson. 2009. "How Much Does Constituency Opinion Affect Legislators' Votes? Results from a Field Experiment." Unpublished manuscript, Institution for Social and Policy Studies at Yale University.

Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54: 297–312.

Clingingsmith, David, Asim Ijaz Khwaja, and Michael Kremer. 2009. "Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering." *Quarterly Journal of Economics* 124: 1133–1170.

Druckman, Daniel. 1994. "Determinants of Compromising Behavior in Negotiation: A Meta-Analysis." *Journal of Conflict Resolution* 38: 507–56.

Druckman, James N. 2003. "The Power of Television Images: The First Kennedy-Nixon Debate Revisited." *Journal of Politics* 65: 559–71.

Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98: 671–86.

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

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

Gerber, Alan S., Donald P. Green, Edward H. Kaplan, and Holger L. Kern. 2010. "Baseline, Placebo, and Treatment: Efficient Estimation

for Three-Group Experiments." *Political Analysis* 18: 297–315.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945–60.

Hyde, Susan D. 2009. "The Causes and Consequences of Internationally Monitored Elections." Unpublished manuscript, Yale University.

Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: The University of Chicago Press.

Iyengar, Shanto, Mark D. Peters, and Donald R. Kinder. 1982. "Experimental Demonstrations of the 'Not-So-Minimal' Consequences of Television News Programs." *American Political Science Review* 76: 848–58.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the New South." *Journal of Politics* 59: 323–49.

Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. "The Effects of Negative Political Advertisements: A Meta-Analytic Review." *American Political Science Review* 93: 851–75.

Loewen, Peter John, Royce Koop, and James H. Fowler. 2009. "The Power to Propose: A Natural Experiment in Politics." Unpublished paper, University of British Columbia.

Lyall, Jason. 2009. "Does Indiscriminant Violence Incite Insurgent Attacks?" *Journal of Conflict Resolution* 53: 331–62.

McKelvey, Richard D., and Thomas R. Palfrey. 1992. "An Experimental Study of the Centipede Game." *Econometrica* 4: 803–36.

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

Mullainathan, Sendhil, Ebonya Washington, and Julia R. Azari. 2010. "The Impact of Electoral Debate on Public Opinions: An Experimental Investigation of the 2005 New York City Mayoral Election." In *Political Representation*, eds. Ian Shapiro, Susan Stokes, Elizabeth Wood, and Alexander S. Kirshner. New York: Cambridge University Press, 324–42.

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles." *Statistical Science* 5: 465–80.

Nickerson, David W. 2008. "Is Voting Contagious?: Evidence from Two Field Experiments." *American Political Science Review* 102: 49–57.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review* 86: 404–17.

Panagopoulos, Costas, and Donald P. Green. 2008. "Field Experiments Testing the Impact of Radio Advertisements on Electoral Competition." *American Journal of Political Science* 52: 156–68.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688–701.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sniderman, Paul M., and Thomas Piazza. 1995. *The Scar of Race*. Cambridge, MA: Harvard University Press.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61: 821–40.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior." *World Politics* 55: 399–422.