

## Observational Studies

### 1. Learning Objectives

After reviewing this chapter readers should be able to:

- Recognize that controlled experimentation (the ability to systematically manipulate variable of interest) is often not possible, and understand that alternative observational study designs are possible but that they entail many potential pitfalls;
- Describes the major threats to the integrity of observational research results such as threats to validity, reliability, statistical inference and generalizability;
- Outlines some ways to improve each step in the research process, including choosing the most appropriate research design, obtaining meaningful measurements, conducting sound statistical analyses and creating adequate standards for the reporting of study findings;
- Understand the many challenges to all observational studies and the need for transparent reporting of any aspect of study design, data collection, or analysis that could materially affect study findings and their generalizability.



## 2. Introduction

Observational studies are ubiquitous, and yet, they are not clearly defined. A classic book on the topic explains that observational studies have two characteristics (Cochran 1983).

1. An objective to study the casual effects of certain agents, procedures, treatments or programs.
2. The investigator cannot use controlled experimentation, for one reason or another. That is, the investigator cannot impose on a subject, or withhold from a subject, a procedure or treatment whose effects he desires to discover, or cannot assign subjects at random to different procedures.

**It is important to appreciate that the intervention is a manipulable alteration in the status quo and, in that important sense, observational studies are akin to experiments.**

And as Rosenbaum (2002: 1-2) observes, "A study without a treatment is neither an experiment nor an observational study. Most public opinion polls, most forecasting efforts, most studies of fairness and discrimination, and many of other important empirical studies are neither experiments nor observational studies."

An ambiguity occurs in the fact that control over study design may be less important than the quality of the design itself. In some cases, there are "natural experiments" in which natural or social processes assign subjects to treatments so that a very strong research design follows. The data may then be properly analyzed as an experiment, even a randomized experiment. A well-known example is the draft for the Vietnam war (<http://www.landscaper.net/draft.htm>). Soldiers were drafted by a lottery, which amounts to random assignment to be drafted or not. Researchers could later study the impact of the draft on such things as subsequent earnings (Angrist, 1990).

**This chapter will focus on the causal impact of manipulable interventions that are not assigned to subjects at random. Who has control is a secondary concern.**

One of the most well-known set of observational studies is the "Framingham Heart Study", which included research on risk factors for heart disease (Mamun, 2003). In this study some risk factors, such as smoking, diet and exercise, were not under the control of researchers, but were nonetheless important to understand because they could be the target of interventions.

The goal of the chapter is not to provide a set of iron-clad rules about what makes for a good observational study, or what features of such studies necessarily should be reported. The goal is to establish a burden of proof. If the suggestions to follow are ignored, the burden is on the researcher to make a strong case that some alternative reporting approach is better. And that rationale needs to be provided in an accessible manner for all to see.

### 3. Descriptive Validity

#### Summaries of the Data

Any evaluation requires summaries of the data, with some summaries more useful than others. Indeed, some summaries can be misleading. Perhaps the most well known example is the manner in which outliers can affect measures of central tendency: the mean, median, or mode. The mean is especially vulnerable. But outliers can dramatically affect many other summary statistics such as the standard deviation, the Pearson correlation coefficient, and the regression coefficient.



#### Example 1: Community Development

The research of Galster and Temkin (2004):502 asks, how can one show whether "efforts by government, community development corporations (CDCs), and for-profit developers to revitalize distressed, inner-city neighborhoods make any demonstrable difference? Put differently, can a method be devised for persuasively quantifying the degree to which significant, place-based investments causally contributed to neighborhoods' trajectories, compared to what would have occurred in the absence of interventions."

A central conceptual concern is how to best define the counterfactual. But they also address a number of statistical issues. They favor a pooled cross-section time series design with neighborhoods as the observational units. An important issue is how best to take spatial dependence into account. Other things equal, neighborhoods close by one another will tend to be more alike than neighborhoods farther away. When the authors regress median neighborhood housing prices on a set of predictors, including a binary variable for an intervention, they use the inverse Euclidian distance between neighborhoods to weight their regressions (Galster and Temkin, 2004:516). But is this a good statistical summary of spatial proximity? It assumes that dependence declines linearly with distance and that the decline is smooth. Yet, the quality of neighborhoods can change sharply in just a few blocks, and breaks in neighborhood continuity caused by freeways, parks, and bodies of water can introduce abrupt changes in the degree of dependence.

### 3. Descriptive Validity

The choice of reported summary statistics can also be guided by disciplinary and substantive concerns. For example, when the outcome variable of interest is binary, what sort of summary statistic should be used? If there is a single treatment group and a single comparison group should one rely on the difference between proportions, the risk ratio, or the odds ratio? When the two proportions are very small, some researchers favor the risk ratio or odds ratio.

Analogous issues come up when the outcome is quantitative. Should the income inequality of a county, for instance, be represented by the standard deviation of household income or the Gini index of household income?

Likewise, when is it appropriate to standardize a variable? Should a measure of infant mortality be the raw number of deaths in the first year of life, the proportion of children who die in their first year of life, or the number of deaths per capita? When should variables be reported in their original units and when should they be reported in standard deviation units (z-scores)? Similar issues arise for higher level summaries.

## Summaries of summary statistics

It is common to fit a statistical model to the data. Many researchers report some measure of fit and may even use that measure to help determine which of several models is best. However, there are many measures of fit.

- **R<sup>2</sup>**: Statistical measure of how well a regression line approximates real data points; an  $r^2$  of 1.0 (100%) indicates a perfect fit.
- **R<sup>2</sup> Adjusted for degrees of freedom**: measures the proportion of the variation in the dependent variable accounted for by the explanatory variables. Unlike  $R^2$ , adjusted  $R^2$  allows for the degree of freedom associated with the sums of the squares. Therefore, even though the residual sum of squares decreases or remains the same as new explanatory variables are added, the residual variance does not.
- **Bayesian Information Criterion**: Bayesian Information Criterion (BIC) is a model selection criterion.  $BIC = n \ln (SS_{RES}/n) + k \ln(n)$  where  $SS_{RES}$  is the residual sum of squares,  $n$  is the number of observations, and  $k$  is the number of estimated parameters.
- **Akaike Information Criterion**: a criterion for selecting among nested econometric models. The AIC is a number associated with each model:  
 $AIC = \ln (s_m^2) + 2m/T$
- **Mallows Cp**: method to find adequate models by plotting a special statistic against the number of variables+1.  $C_p = SS_{res}/MS_{res} - N + 2p$ ,

where  $SS_{res}$  is the residual sum of squares for the model with  $p-1$  variables,  $MS_{res}$  is the residual mean square when using all available variables,  $N$  is the number of observations, and  $p$  is the number of variables used for the model plus one.

The choice of a fit measure can matter. Care must be exercised in such a choice.

### 3. Descriptive Validity

In short, it is important to explain the choice of particular statistical summaries relied upon in the analysis. One might say, for instance, that the median was chosen because it is more robust than the mean to outliers, and there was evidence of outliers in the data. Or one might say that a reported regression model was determined by the Bayesian Information Criterion because it was reasonable to assume that there were clear distinctions in principle between regressors that belonged in the model and regressors that did not.

Therefore, among the matters to be reported are:

1. The choice of any summary indices;
2. The choice to standardize or not standardized any variables;
3. The choice of univariate summary statistics;
4. The choice of bivariate or multivariate summary statistics; and
5. The choice of any statistical summaries used in model construction or evaluation.

## 4. External Validity

### Credible Generalizations

The next topic to be considered is where the data come from. Answers to that question will indicate the kinds of generalizations that can be made and the obstacles to be faced making any inferences beyond the data on hand.



#### Example 2: Promoting Education in Youth

When troublesome adolescents are sent to special alternative schools where strict discipline is enforced, does the experience increase or decrease the likelihood of subsequent success in regular public schools? Wolf and Wolf (2008) report the results of a program meant to break the "school to prison pipeline" for a number of school-aged children from the Syracuse (New York) city school district. The program had seven components:

1. Systematic support for students anticipating a transition from special alternative schools to mainstream schools;
2. Out-of-school activities meant to improve social and academic skills;
3. Promotion of "bonding" between the youth and "caring" adults.
4. Counseling and referrals;
5. Support groups for students with incarcerated loved ones.
6. Regular contact with parents; and
7. Collaborative training for teachers, administrators, and other school staff.

The students in the study were essentially a convenience sample of all students in the Syracuse city school district, and Syracuse itself is a convenience sample of large cities in the United States. Yet, the observational study was motivated by the desire to learn about the efficacy of such interventions in general. What would be the point of learning whether the intervention worked for the several hundred students in the study unless that information could inform future policy decisions? In this instance, the intervention did not seem to produce beneficial effects overall and may even have produced some undesirable effects. But what kinds of generalizations can be properly justified?

## 4. External Validity

**The point of undertaking an observational study is usually larger than the data set to be analyzed. Yet, making generalizations from the study is difficult.**

Ideally, the data are generated by probability sampling from a well defined population. Statistical generalizations from a probability sample to the population from which it was drawn can then be a routine matter. In the absence of probability sampling, any credible generalizations depend on strong theory and/or replications.

For the former, one would need widely accepted theory showing that for programs sufficiently "like this" and study subjects sufficiently "like this," one can expect the same kinds of effects. For Example 1, no compelling theory existed that provided requisite specificity. There was apparently no theory that defined clearly the set of young people in alternative schools for whom the intervention was appropriate. For example, given any collection of troubled high school students, it would be difficult to determine who would be appropriate for the program and who would not.

For the latter, there would need to be a substantial number of studies of programs sufficiently "like this" and study subjects sufficiently "like this" so that firm empirical generalizations can be constructed. The results of several earlier studies related to the program were reviewed, but the program evaluated was not a very close replication of past efforts. Moreover, the past studies reviewed did not speak with one voice about what works and for whom.

## 4. External Validity

**Just as important for how the data collection was designed is how well the data collection was implemented.**

Even with a well-defined population and a probability sample from that population, the data actually on hand may make generalization difficult. For example, if the data are from a sample survey, a low response rate can be devastating. If the response rate is low and there is reason to believe that the responders differ in important ways from nonresponders, generalization to the population of responders and nonresponders can be problematic. If the response rate is high or responders and nonresponders seem much alike, one may be able to proceed as usual.

In practice, the amount of nonresponse and the differences between responders and nonresponders are matters of degree. There is, therefore, no threshold above which generalizations are justified and below which they are not. The degree of accuracy needed for the study must be factored in. At one extreme, response rates of 15% are sometimes deemed acceptable in marketing studies. At the other extreme, response rates of over 90% are often required for government studies from which important economic and political decisions are made.

Generalizations from a well-designed observational study also can be undermined by study attrition. When study subjects need to be followed over time, some study subjects may refuse to cooperate or otherwise not be found. The issues are much the same as for the response rate. The two questions are how many subjects are lost and how different are they from the subjects who cooperate for the life of the study. In the Wolf and Wolf (2008) study, for example, there was attrition because some students' families moved out of the study area or did not complete the program. The problem was addressed by making attrition one of several jointly estimated outcomes.

The distinction between nonresponse and attrition can get fuzzy in practice. If the data are collected after the intervention has been delivered, nonresponse can have some of the same consequences as attrition. Subjects are not lost from the initial pool of subjects, but the

nonresponse may be related to the intervention. This is a concern normally raised about attrition: estimates of any intervention effects may be biased because those subjects who are lost from the intervention group may differ in important ways from those lost from the alternative condition. However, this is less an issue of awed generalization and more an issue of potentially biased treatment effect estimates. We return to this topic below.

## 4. External Validity

### Creating a Research Report

What then should be reported? Each entry in the following list should be included in any research report unless it is not relevant to the study.

1. For probability samples
  - a. What is the population?
  - b. What was the sampling design?
  - c. Why was the sampling design chosen (e.g., feasibility, power calculations, etc.)?
  - d. What is the definition of the response rate, how was it computed, and what is the figure?
  - e. How do responders tend to differ from nonresponders?
  - f. What proportion of the study subjects was lost through attrition?
  - g. If there is no attrition, that too should be reported.
  - h. How do those lost through attrition tend to differ from those not lost through attrition?
  
2. For nonprobability samples or populations
  - a. What are the replications that are being used to support any generalizations?
  - b. How does existing theory support any generalizations?

In summary, there are three methods by which generalizations can be made:

1. Probability sampling from a well-defined population;
2. Sound and widely accepted theory; and
3. Sound replications directly relevant to the study on hand that speak with one voice about the findings.

If the data were generated by probability sampling, it is important to describe the results of any power analyses done as the sampling design was constructed, the sampling design arrived at, and how well that design was implemented. Thus, for example, low response rates and sample

attrition during the life of the study can decimate the best of sampling designs and should be reported. Note also that post-hoc power analyses are not the same thing as power analyses done before the data were collected and are generally a bad idea in any case (Hoenig and Heisey, 2001).

If the data were not generated by probability sampling, one must rely on theory and/or replications. In both instances, more should be provided than a list of citations. A strong rationale should be written for any generalizations that are made. This is often difficult to do in program evaluation because the extant theory can be very weak and true replications rare.

## 5. Construct Validity

### Measurement

There is no argument that all variables important to an analysis should be well measured. There is also no argument that most variables will be measured in an imperfect manner. Therefore, two questions need to be addressed:

1. Are the requisite variables included in the dataset? and
2. How well are they measured?

As explained more fully later, it is as important to have well-measured response variables as it is to have good measures of the intervention(s) and comparison conditions, along with good measures of all confounders.

**We focus, therefore, on measurement quality. Sometimes measurement quality is called construct validity.**

To take a recent example, there is growing interest in measuring how effective colleges are in educating undergraduates. One possible measurement tool is the "College Learning Assessment" (CLA) instrument. "The CLA focuses on the institution (rather than the student) as the unit of analysis. Its goal is to provide a summative assessment of the value-added by the school's instructional and other programs (taken as a whole) with respect to certain important learning outcomes" (Klein et al., 2007:418). The natural question is how well the CLA achieves these aims.

## 5. Construct Validity

To take another example, research on the impact of early child care and educational interventions requires sensible measures of those activities. As Layzer and Goodson (2006):556 note "There is a widespread belief that high-quality early care and education can improve children's school readiness. However, debate continues about the essential elements of high-quality experience, about whether quality means the same things across different types of care settings, about how to measure quality, and about the level of quality that might make a meaningful difference in the outcomes of children."

In their article they address four questions:

1. How is the quality of child care environment commonly defined and measured?
2. Do the most commonly used measures capture the child's experience?
3. Do measures work well across all care settings?
4. Are researchers drawing the correct conclusions from studies of child care environments and child outcomes?

**Good measurement can be boiled down to two features: validity and reliability. Both in practice are matters of degree. For validity the issue is how well you are measuring what you think you are measuring.**



### Example 3: Recovery Management

Substance abuse is often a chronic problem for which several interventions are needed so that each responds to where in the life course an individual falls. One implication is a shift from an acute care paradigm to a chronic care paradigm. Rush and his colleagues report on the results of an intervention called "Recovery Management Checkups" (RMC) designed to help "people with substance abuse disorders by level of co-occurring mental disorders..." (Rush et al., 2008:7). "The RMC intervention targets individuals who have previously participated in treatment and are now living in the community using substances. The intervention ... aims to provide immediate linkage back to substance abuse treatment on the basis of need, thus expediting the recovery process. Key components include, for example, assessing eligibility for the intervention and need of treatment, transferring participants in need of treatment from the interviewer to a linkage manager for a brief intervention, linking participants to the intake assessment, and ultimately linking participants to treatment" (Rush et al., 2008:8). A key measurement issue is to determine who is in need of treatment. For this study, such a person was defined as a study participant living in the community (vs. incarcerated or in treatment) who was not already in treatment and answered yes to any of the following questions:

1. During the past 90 days, have you used alcohol, marijuana, cocaine, or other drugs on 13 or more days?
2. During the past 90 days, have you gotten drunk or been high for most of 1 or more days?
3. During the past 90 days, has your alcohol or drug use caused you not to meet your responsibilities at work/school/home on 1 or more days?
4. During the past month, has your substance use caused you any problems?
5. During the past week, have you had withdrawal symptoms when you tried to stop, cut down, or control your use?
6. Do you feel that you need to return to treatment?"

The alpha (measuring internal consistency) reported for these items was .85.

## 6. Measurement Validity

**Measurement validity begins with an assessment of how the target of the measurement is conceptualized. One problem can be "over-coverage." Another problem can be "under-coverage."**

In the case of IQ, if the concept of general cognitive ability does not include a full range of cognitive skills (e.g., musical), there is under-coverage. If the concept of cognitive ability includes attributes that are really culturally based (e.g., vocabulary), there is over-coverage.

In the case study by Rush and his colleagues (2008), the questionnaire items seem intuitively sensible, but no definition of the need for treatment is provided. Consequently, both under-coverage and over-coverage are possible. For example, all but one of the questionnaire items provide a reference time period (e.g., 90 days). There are good reasons for this from the point of view of item construction. One needs to specify a suitable interval for recall. But are there people in need who will be missed without a longer recall interval? And are there features of need that are not addressed with items (a)-(e), or items that do not really reflect what might be meant by "need?" If yes, the measure is likely to be systematically inaccurate. Some refer to this as measurement bias.

## 6. Measurement Validity

A measurement with systematic error is technically not considered a "valid" measurement although we tolerate in practice small amounts of systematic error. Thus, measurement validity is a matter of degree. However, it is difficult to know how large the bias really is. If we knew the direction and size of the bias, we could correct for it.

Sometimes a measure is called a "proxy." For example, the number of homicides in a city could be a proxy for the seriousness of the city's crime problems. Clearly, this proxy risks under-coverage. A proxy measure by definition has systematic error. Calling such a measure a proxy does not make it sound. Implicitly, however, the measure is taken to be good enough to be useful. In practice, whether a proxy measure is really good enough needs to be argued.

**Beyond conceptual issues, there can be operational problems in the steps by which the measurement is done. The "recipe" by which the concepts to be measured are translated into the activities of these steps is called an "operationalization."**

Thus one speaks of "operationalizing" the concept. Faulty operationalization also leads to systematic measurement error. For example, in the criminal justice area, there can be operational errors if one relies on data recorded by patrol officers when they fill out offense forms. If the explanations given to police officers about how to fill out an offense form are wrong, there will be operationalization errors. The police may do as they were told, but what they were told is wrong. They may, for instance, be given unclear guidance about what information to include about the victim. For the drug treatment case study, a lot would depend on whether operationalizing, say, drug problems without making distinctions between different kinds of drugs may misrepresent how the need is characterized.

## 6. Measurement Validity

Sometimes, of course, the problem is with how the operationalizations are implemented. In the police illustration, patrol officers will sometimes write down the facts incorrectly even if the instructions are clear. For example, the narrative may indicate that there was a forced entry into a warehouse, when actually an employee just failed to properly secure a window. For the case study described above, an obvious operational problem could be the accuracy of respondents' recall.

In short, even well-defined measures can have validity problems. The translation of the concept into concrete procedures can be flawed and/or the ways the measurement procedures are carried out can be flawed. Both can lead to serious systematic errors.

One must also be careful about reification in which an operationalization is taken to be the real thing. A good example is when IQ tests (discussed earlier) are taken to be intelligence itself. The same applies to SAT scores if they are taken as academic ability itself. In the case study, the survey-based measure of need might be taken as need itself. Reification can lead to serious misunderstandings in part because interventions can be directed to changing the measure instead of what it is supposed to measure. Perhaps the best example is when teachers teach to the test under provisions of No Child Left Behind.

## 6. Measurement Validity

How does one get a handle on measurement validity? The usual strategy is to do special studies in which the true values are obtained and then compared to the measured values. For example, one can ask in a survey how much money people gave to a particular charity (which is likely to be over-reported) and then check the charity's records to see if the reported value is correct. One can do the same thing, at least in principle, for drunk driving arrests, which tend to be underreported. For the case study, the gold standard might be true clinical assessments of need.

## 7. Measurement Reliability

What some people call "noise", also called "chance error," does not create systematic error. It makes a measure unreliable. If one measured the same thing repeatedly, but as much as possible in the same manner, the results will likely vary, at least a bit. However, the mean of the measures could be a good approximation of the "true" value. The noise cancels out in a large number of measures. For example, the urine drug tests given to individuals on parole are generally thought to be usefully valid. But the measures have some "wiggle" in them. Even measures for the same person on the same day will likely differ at least a bit from one another. But the average over many tests could be a good approximation of the noise-free value.

A key complication in practice is that for most measures we use, there is only a single measurement, and that measurement is likely to be inaccurate by some (unknown) chance amount that is not cancelled out. Ideally, the variation across units being measured is not being dominated by noise.

**One way to get some handle on this is to determine whether the measure varies in sensible ways with other measures to which it should be related.**

If the chance components for each measure are approximately independent of one another, this can be a very helpful analysis. For example, city neighborhoods with a lower median household income should have more crime, more young people dropping out of school, and higher infant mortality rates.

This idea sometimes can be exploited more directly to estimate the reliability of a given measurement procedure. For example, it is common to break up some multiple item instrument, such as a measure of depression, into two sets of randomly chosen items. The correlation between the two "parallel" sets of items is an estimate of the reliability of the instrument overall. The higher the correlation, the more reliable the instrument.

## 8. Formal Representation

For any given observational unit  $i$ , let

$$(1) x_i = T_i + \beta_i + \varepsilon_i,$$

where  $x_i$  is the measurement,  $T_i$  is the "true" value,  $\beta_i$  is the bias, and  $\varepsilon_i$  is the chance error.

The  $\varepsilon_i$  is assumed to have been generated at random from a distribution with a mean of zero; on the average, the noise cancels out. It is also, under this simple model, unrelated to  $T_i$  or  $\beta_i$ . For example, larger values of  $\varepsilon_i$  are not more likely when  $T_i$  or  $\beta_i$  are larger. In short, if  $\beta_i$  is not zero, especially if  $\beta_i$  is large, there can be substantial bias. There are also problems if  $\varepsilon_i$  is large. Then, reliability will tend to be low. Ideally,  $\beta_i$  and  $\varepsilon_i$  should be small.

There are additional problems if  $\varepsilon_i$  is related to  $\beta_i$  or  $T_i$ . For example, if the size of the bias is related to the size of the "true" value, it can be difficult to obtain a good fix on the bias. Thus, in areas where there is a lot of crime, people may be less likely to report it. They may believe there is no point or that there could be retaliation. One result is that the underreporting of crime can be higher in high crime neighborhoods. Therefore, the bias is not constant. The size of  $\beta_i$  depends on the size of  $T_i$ . This is a major complication not captured by the simple equation above.

For an observational study the following information should be reported for each measure:

1. The definition of what is being measured
2. A formal representation of how the measurement process is assumed to function
3. Justification for that representation
4. Any information on validity
5. Any information of reliability

## 9. Internal Validity

### Causal Inference: Internal Validity

Modern discussions of causal inference are based on how statisticians have come to think about cause and effect. The statistical framework was developed by Neyman in 1923 and later extended by Rubin (1974) and Holland (1986). It is sometimes called the "Rubin Causal Model."



#### Example 4: Causal Effect Defined by Rubin

Rubin defines a causal effect:

Intuitively, the causal effect of one treatment,  $E$ , over another,  $C$ , for a particular unit and an interval of time from  $t_1$  to  $t_2$  is the difference between what would have happened at time  $t_2$  if the unit had been exposed to  $E$  initiated at  $t_1$  and what would have happened at  $t_2$  if the unit had been exposed to  $C$  initiated at  $t_1$ : 'If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone,' or because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone.' Our definition of the causal effect of the  $E$  versus  $C$  treatment will reflect this intuitive meaning.

According to the RCM, the causal effect of your taking or not taking aspirin one hour ago is the difference between how your head would have felt in case 1 (taking the aspirin) and case 2 (not taking the aspirin). If your headache would remain without aspirin but disappear if you took aspirin, then the causal effect of taking aspirin is headache relief (Rubin, 1974:689).

There are observational units: people, neighborhoods, police departments, schools, business establishments or other entities. In the simplest case, there is a binary intervention. Some of the units are exposed to an intervention, and the other units are exposed to an alternative. The intent is to estimate the intervention's causal effect. The classic examples come from research

on the impact of job training, housing vouchers, or instructional innovations in schools. But equally interesting studies use larger observational units. For example, one might want to learn how sanctions applied to employers who hire undocumented workers affect the flow of immigrants into and out of a given state. Or, one could examine the possible impact on water conservation of a city's educational campaigns during a drought.

## 9. Internal Validity



### Example 5: Firearms

Domestic violence exacts a well-documented and costly toll on victims and their families. In the United States "on the average, 3.5 people are killed by intimate partners every day, and many others are injured" (Vigdor and Mercy, 2006:313). An important question, therefore, is whether laws restricting access to firearms for individuals with a history of domestic violence can reduce intimate partner homicides (IPH). Vigdor and Mercy try answer this question using states as the observational units.

"The states that have laws limiting access to guns by abusers passed their legislation at different times. We exploit this time variation by effectively comparing IPH rates before and after passage of the law in states that enacted these laws with those in states that did not pass such a law. Although we cannot be certain that we are isolating the impact of the laws, the time variation in the effective date of the laws reduces the likelihood that we are capturing the effect of an omitted shock affecting all IPH rates."

Fully appreciating states can vary on the other factors that can affect IPH, they use a negative binomial regression model that includes a large number of covariates. They find that restraining- order laws help to keep perpetrators and victims apart and reduce IPH. Other kinds of interventions, such as gun confiscation laws, had no demonstrable impact.

Each observational unit is assumed to have two potential outcomes. There is an outcome if the unit is exposed to the intervention. There is another outcome if the unit is exposed to the alternative. These outcomes can vary across units and are hypothetical. Using the Vigdor and Mercy case study, a given state has a potential number of IPHs if a law is passed restricting the access of batterers to firearms, and a potential number of IPHs if there is no such law.

Suppose we let  $Y_i(1)$  denote the hypothetical IPH count if state  $i$  enacts the relevant legislation, and  $Y_i(0)$  denote the hypothetical IPH count if state  $i$  does not enact that legislation. The causal effect of the legislation can be defined as  $[Y_i(1)-Y_i(0)]$ , although occasionally  $[Y_i(1)/Y_i(0)]$  is used instead.

It is impossible to observe both  $Y_i(1)$  and  $Y_i(0)$ . Either a given state passes the relevant legislation or it does not. Suppose we let  $W_i = 1$  if state  $i$  enacts the legislation, and  $W_i = 0$  if state  $i$  does not enact the legislation.

Then the observed outcome for a given state is:

$$(2) Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1)$$

$Y_i$  and  $W_i$  can be observed. But there is no way to map what can be observed back to the definition of a causal effect for a given state. Either the legislation passes or it does not. Consequently, we shift to group comparisons. In this example, attention is directed to the average IPH count of the states that passed the relevant legislation compared to average IPH count of the states that did not pass the relevant legislation. The difference between the two is an estimate of the average treatment effect (ATE).

## 9. Internal Validity

**ATE: The average treatment effect is not the only kind of causal effect that can be estimated (Imbens, 2004).**

- One can distinguish between an estimate of the average causal effect in some well-defined population and the estimate of the average causal effect solely for the set of study subjects.

**Question:** What would happen if the full pool of study subjects consisted of only the subset of subjects who were actually treated?

- One can also distinguish between the estimated average causal effect on the treated and the estimated average causal effect on the untreated.

**Question:** What would happen if the full pool of study subjects consisted of the subset of subjects who were actually not treated?

**These various definitions go to the question of external validity:** to what pool of subjects do the estimated average treatment effects apply?

**All of the definitions of what is to be estimated introduce a key problem:** how does one separate the impact of the intervention from the impact of other factors that could affect the outcome?

For example, the states that passed the IPH legislation may differ systematically from those that did not in ways related to their potential outcomes. The treatment states may be more likely to have higher rates of intimate partner homicide due to a larger fraction of households living below the poverty line. Unless this is taken into account, an estimate of the average treatment effect risks being biased against any beneficial impact of the legislation.

**There are three generic strategies to address this problem (Imbens, 2004).**

- **Stratification**
- **Weighting**
- **Regression**

## 9. Internal Validity

### Stratification

First, one can stratify by confounders, although one usually needs access to a large sample. The basic idea is to group observations into different strata according to their values on variables that are likely to be related to  $W_i$  and potential outcomes. That is, within a stratum, observations are made homogenous with respect to the confounders.

For example, a study of the impact of housing vouchers on whether a family moved would probably have to stratify variables such as household income, the number of school aged children in the household, and the current rent the household is paying. However, unless there are very few confounders and unless they are largely categorical, it is very difficult to achieve complete homogeneity within each stratum. One rapidly runs out of cases. An alternative is to use a form of "caliper matching" in which approximate homogeneity is the goal. A key diagnostic is whether after approximate matching the sample is balanced on the covariates. Do the different treatment groups have effectively the same distributions on all of the confounders?

When satisfactory homogeneity within strata is achieved, one simply computes in each the difference between the mean of the treatment group and the mean of the group receiving the alternative. The average treatment effect is then just the weighted average of these within-stratum average treatment effects. The weights are determined by the within-stratum sample sizes.

## 9. Internal Validity

### Weighting

When there are a large number of confounders, and especially if they are not categorical, another approach is to reduce the size of the predictor space through propensity scores. One builds a statistical model from which is constructed an estimated probability of assignment to each treatment group and comparison. These probabilities are called "propensity scores." Then one can stratify by ranges of propensity score values alone. If the statistical model from which the propensity scores are built is effectively correct, one can achieve very much the same results as approximate stratification (Rosenbaum, 2002). It is also possible to use propensity scores as weights when computing outcome contrasts between the treatment and comparison groups, but such weighting can introduce serious difficulties even when it is formally appropriate (Freedman and Berk, 2008).

Beneath the stratification approach is a very important assumption. Conditional on the covariates, one imagines that nature undertakes the equivalent of a randomized experiment. The trick is to condition properly on the right variables. With that done, one can proceed as if one had random assignment to treatment and comparison groups within strata. It is impossible to definitively determine if the imagined underlying experiment corresponds well to the reality but sensitivity analyses can help (Rosenbaum, 2002: Chapter 4). The basic idea is to determine empirically if introducing different amounts of selection bias into the treatment and comparison groups would change the study's overall conclusions about causal effects. If it would take an implausibly large amount of bias to alter the study's conclusions in an important way, then the results acquire additional credibility, and the underlying randomized experiment gains credence as well.

## 9. Internal Validity

### Regression

Another approach is to model the response as a function of the intervention and the confounders. Regression models of various kinds are popular, especially among economic and sociological researchers. The many problems with regression modeling are now articulated in any number of sources (Berk, 2003; Freedman, 2005a; Morgan and Winship, 2007). We need not review those problems here. Suffice to say, one's results are only as good as the model. And the models are too often unconvincing. Just as for the stratification strategy, there is an underlying conception of how the data were generated. There is no longer a hidden randomized experiment conditional on the covariates. Rather, there is a causal model representing how the response variable values are determined. Whereas the stratification approach focuses on the manner in which observational units wind up in the treatment group or the comparison group, the causal modeling approach focuses on factors, including the intervention, that affect the response. These are very different visions that often need to be analyzed differently (Freedman, 2008a).

The diagnostics one can apply differ as well. Sensitivity analyses are difficult to formulate. Rather, attention is directed to properties of the residuals as stand-ins for the random disturbances in the model. There is also concern about the functional forms by which the regressors are related to the response. An excellent introduction to regression diagnostics can be found in the textbook by Cook and Weisberg (1999). Some appropriate caveats are provided by Freedman (2008b).

## 9. Internal Validity

The stratification, weighting, and regression approaches all depend on having the correct set of well-measured predictors.

**Each approach can in principle be strengthened by the design through which the data were generated.**

For example, if the group exposed to the intervention and the group exposed to the alternative are very similar to begin with, less importance may be attached to how the confounders are handled. Thus, in a study of the impact of housing vouchers, households that received vouchers might be compared to households from the same neighborhood that did not. A conventional before-after study is often strengthened if there is a time series of observations before the intervention and a time series of observations after: an interrupted time series design. For example, the impact in a single city of a water conservation campaign might lead to a more credible analysis if there are many months of water consumption figures before the campaign and after the campaign.

**Perhaps the strongest of such quasi-experimental designs depends on assignment through a covariate.**

Suppose, for example, prison inmates are assigned to either a high or low security setting depending on a risk assessment score computed upon admission. Inmates who score above a certain value are placed in high security surroundings, and those who score at or below that value are placed in low security surroundings. A regression analysis for some outcome, such as misconduct in prison, that uses only the risk assessment score and the security level that resulted can in principle produce an unbiased estimate of the impact of the placement (Berk and de Leeuw, 1999). This is an example of a regression discontinuity design, which is often the most effective quasi-experimental design available. Clearly, the regression analysis applied is enormously strengthened.

**There can also be features that appear when the design unfolds in practice that signal significant problems.**

In particular, differential attribution from the treatment and comparison groups can lead to serious biases.

For example, if for a welfare-to-work program the treatment is seen to be far more burdensome than the alternative, less motivated individuals will be more likely to be lost through attrition in the treatment group. An estimate of any beneficial effects from the treatment might then be biased upward; individuals with worse job prospects are more likely to be lost to the experimental condition.

**An equally troubling prospect is possible failure to deliver the intervention and control conditions to all of the appropriate study subjects.**

For example, an intervention built on tutoring school children who are not performing well academically will stumble if the children do not show up for the extra help. A failure to deliver the intervention and/or the alternative as designed raises the question of what should be analyzed: the intervention as designed or the intervention as delivered. The former means little more than proceeding as usual while being careful about how the estimated causal effects are interpreted. This is sometimes called an "intention-to-treat analysis." It is common, for instance, to find biases toward zero for any estimated treatment effects. The latter, however, implies that a rather more complicated set of analyses need to be undertaken. In addition to the usual concerns about selection into the treatment and comparison groups, one must adjust for the processes by which delivery of the treatment and comparison condition fails. The issues can get rather technical. For a good discussion in the context of randomized designs see Freedman (2005b). The lessons carry over to the analysis of observational data.

## 9. Internal Validity

Sometimes discussions of the credibility of causal inferences are undertaken under the rubric of "internal validity," and the conclusions are always a matter of degree. The following need to be addressed in a report on an evaluation.

1. What is the kind of causal effect(s) that is being estimated (e.g., average treatment effect on the treated)?
2. What is the research design (e.g., nonequivalent control group design)?
3. How well was the design implemented (e.g., was there serious attrition)?
4. What were the statistical approaches used for the causal effect analysis?
5. What are the vulnerabilities of those approaches?
6. What diagnostic procedures were used and what did they show?

## 10. Statistical Conclusion Validity

Most data are generated in part by chance processes. It is impossible to know in advance exactly what the data will show, and whether, if the data were generated a second time, they will differ at least a bit in ways that can not be anticipated. This uncertainty in the data means that any summary statistics computed and any conclusions that follow will be subject to uncertainty as well.



### Example 6: Survey Interviews in Inner City

Conducting survey interviews can be especially problematic in inner city areas. Residents may be suspicious of outsiders and even hostile if they suspect that they are just being used to advance some purely academic agenda. Holbrook and her colleagues (2006) studied the use of indigenous interviewers in such circumstances. A key issue was the quality of the data compared to data collected by professional interviewers. To this end, they compared summary statistics from survey interviews conducted by local residents to those from survey interviews conducted by professional interviewers. They also built regression models to characterize how the backgrounds of interviewers might affect respondent answers to sensitive questions. The authors do not discuss how they conceptualized the sources of uncertainty. Well over 70 hypothesis tests were conducted. No discounting of p-values is mentioned. Moreover, p-values are typically not reported. Rather one is given the popular "starsystem:" one star for  $p < .10$ , two stars for  $p < .05$ , and three stars for  $p < .01$ .

## 10. Statistical Conclusion Validity

There are three generic ways to think about how uncertainty in data is introduced.

**First, there is a well-defined population with its variables treated as fixed.** If one computes, for example, the mean of each variable in the population many times, each variable's mean would always be the same. The data on hand are a probability sample from the population. Because the composition of each sample is the product of a chance mechanism, it cannot be exactly anticipated and the composition will change from sample to sample in unpredictable ways. Although in the population the variables are treated as fixed, the variables when sampled become random variables. Sometimes this process is called design-based sampling; the data are generated through an explicit sampling design. Most government surveys and political polls fall within this framework.

**A second approach is called model-based sampling.** The data on hand are seen as a realization of a chance process, usually a natural one. That process is then characterized by a statistical model. Linear regression models are a popular example. Suppose, for example, we let

$$(3) Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \varepsilon_i,$$

where  $Y_i$  is the response variable,  $W_i$  is a binary treatment indicator variable,  $X_i$  is a confounder, and  $\varepsilon_i$  is a random perturbation. The predictors  $W_i$  and  $X_i$  are usually treated as fixed. So,  $Y_i$ , as a linear combination of the two, would be fixed as well were it not for  $\varepsilon_i$ . The model assumes that  $\varepsilon_i$  is a realization for observation  $i$  of a random variable with a mean of zero, constant variance, uncorrelated with realized perturbations of other observations, and with  $W_i$  and  $X_i$ . Equation 3 is a theory of how natural processes generated the data on hand. In this case, nature is assumed to generate the data by using a linear model.

**A third approach can be seen as another form of model-based sampling.**

As discussed under causal inference, observational studies can be viewed as randomized experiments undertaken by nature, conditional on a set of covariates.

One can consider the set of study subjects, conditional on the covariates, assigned to treatment and comparison conditions by random assignment. There is no uncertainty in the potential response of any given observational unit. But there is uncertainty in the observed response because that response depends on the value of  $W_i$ , which nature makes a random variable. This is variant of model-based sampling because the account of how nature generates the data is a model and because the process can be seen as sampling. Conditional on the covariates, nature samples some of the observations at random and assigns them to the treatment group. The remaining observations are placed in the comparison group.

## 10. Statistical Conclusion Validity

**It cannot be overemphasized that to be useful, each account of how the data are generated has to be a good approximation of what really happened.**

The account derived from design-based sampling must correspond to how the probability sampling plan was implemented in the field. Information is often available to help make this assessment (e.g., reported response rates). Model-based accounts can be far more difficult to evaluate because thorough and accurate information on how nature actually proceeded is required. Sometimes there is useful information available. For example, interviews with experimental and control subjects may provide insights about why some individuals chose the intervention and why other individuals chose the comparison condition. But too often the information one would need to make confident assessments is not available. Then researchers often fall back on their disciplinary theory, which is too often insufficiently precise or convincing.

Given a strong case for the particular chance mechanism by which uncertainty in the data has been introduced, confidence intervals and formal statistical tests can follow. The tests need to be explicitly formulated before the data are examined. Hypotheses that are stated after a look at the data undermine the p-values that follow. Generally, the p-values will be too small; there is false power. Another common error is a failure to discount p-values after multiple statistical tests. The problem is that the researcher is capitalizing on chance. For example, one in twenty tests will on the average be "statistically significant" at the .05 level when the null hypothesis is true. There are many interesting proposals for how to constrain this "false discover rate," the details of which are beyond the scope of this discussion (See, for example, Benjamini and Hochberg, 1995).

## 10. Statistical Conclusion Validity

**A deeper problem is the use of statistical tests after model selection procedures are applied.**

Basically, the tests will have unknown properties and cannot be relied upon; the model winnowing process invalidates the tests. As Leeb and Pötscher (2006): 2554 observe,

"...a post-model-selection estimator here refers to the combined procedure resulting from first selecting a model (e.g., by a model selection criterion such as AIC or by a hypothesis testing procedure) and then estimating the parameters of the selected model (e.g., by least-squares or maximum likelihood), all based in the same data. We show that it is impossible to estimate this distribution with reasonable accuracy, even asymptotically."

The best response to this problem is to have a training data set with which to build the statistical model and a test data set with which to undertake any statistical inference. Ideally the two data sets would be random samples from the same population or random realizations of the same data generating process. If one has a large enough data set on hand, an equally good strategy is to randomly divide the data into two parts and treat one part as a training sample and the other part as a test sample.

Putting all this together leads to the following reporting suggestions. One should report:

1. The account being used to characterize the uncertainty (e.g., design-based sampling);
2. The names of any tests used;
3. The null and alternative hypotheses for any statistical tests;
4. Any distributional assumption being made and their credibility for the data on hand;
5. The actual p-values of any statistical tests;
6. The degrees of freedom for any statistical tests;
7. Any model selection procedures used before the reported tests; and
8. The methods used to adjust for multiple tests.

## 11. Summary

It is difficult to draw convincing conclusions from observational studies. There are many potential pitfalls and many research decisions for which clear methodological advice does not exist. The reporting burdens are, therefore, far heavier than for studies that can proceed largely by well-accepted recipes. Beyond the reporting suggestions discussed above, there will often be additional matters of a study-specific nature that will need to be disclosed. Good practice depends on reporting anything about the data collection and analysis that could materially affect the findings.

## 12. References

- Angrist, J. (1990). "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security records." *The American Economics Review*. 83(3): 313-336.
- Benjamini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society*. (Series B) 57(1): 289-300.
- Berk, R.A. (2003). *Regression analysis: A constructive critique*. Sage Publications, Newbury Park, CA.
- Berk, R.A., and de Leeuw, J. (1999). "An evaluation of California's inmate classification system using a generalized regression discontinuity design." *Journal of the American Statistical Association*. 94(448): 1045-1052.
- Cochran, W.G., (1983). *Planning & analysis of observational studies*. New York: John Wiley and Sons.
- Cook, R.D. and Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: John Wiley and Sons.
- Freedman, D.A. (2005a). *Statistical models: Theory and practice*. Cambridge University Press, Cambridge.
- Freedman, D.A. (2005b). "Statistical models for causation: What inferential leverage do they provide?" *Evaluation Review*. 30(6): 691-713.
- Freedman, D.A. (2008a) "On regression adjustments to experimental data." *Advances in Applied Mathematics*. 40(2): 180-193.
- Freedman, D.A. (2008b). "Diagnostics cannot have much power against general alternatives." Working paper at Berkeley.
- Freedman, D.A., and Berk, R.A. (2008). "Weighting regressions by propensity scores." *Evaluation Review*. 32(4): 392-409.

- Galster, G., Temkin, K., Walker, C., and Sawyer, N. (2004). "Measuring the impacts of community development initiatives: A new application of the adjusted interrupted time-series method." *Evaluation Review*. 28(6): 502-538.
- Hoenig, J.M. and Heisey, D.M. (2001). "The abuse of power: The pervasive fallacy of power calculation for data analysis." *The American Statistician*. 55: 19-24.
- Holland, P. (1986). "Statistics and causal inference." *Journal of the American Statistical Association* 8: 945-60.
- Imbens, G., (2004). "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and Statistics*. 86: 4-30.
- Klein, S., Benjamin, R., and Bolus, R. (2007). "The collegiate learning assessment: facts and fantasies." *Evaluation Review*. 31(5): 415-439.
- Layzer, J.I., and Goodson, B.D. (2006). "The 'Quality' of Early Care and Education Settings" *Evaluation Review*. 30(5): 556-576.
- Leeb, H., and Potscher, B.M. (2006). "Can one Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics*. 34(5): 2554-2591.
- Mamun, A. (2003). *Life history of cardiovascular disease and Its risk factors - multistate life table approach and application to the framingham heart study*. Amsterdam: Rozenberg Publishers.
- Morgan, S.L., and Winship, C. (2007). *Counterfactuals and causal inference: Methods and principle for social research*. Cambridge University Press, Cambridge.
- Neyman, Jerzy. 1923 [1990]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles." *Statistical Science*. 5 (4): 465–472. Translation by Dorota M. Dabrowska and Terence P. Speed.
- Rosenbaum, P.R. (2002). *Observational studies, Second Edition*. New York: Springer.
- Rubin, D. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*. 66: 688-701.

Rush, B.R., Dennis, M.L., Scott, C.K., Castel, S., and Funk, R.R. (2008). "The interaction of co-occurring mental disorders and recovery management checkups on substance abuse treatment participation and recovery." *Evaluation Review*. 32(1): 7-38.

Vigdor, E.R., and Mercy, J.A. (2006). "Do laws restricting access to firearms by domestic violence offenders prevent intimate partner homicide." *Evaluation Review* .30(3): 313-346.

Wolf, E.M., and Wolf, D.A. (2008). "Mixed results in a transitional planning program for alternative school studies." *Evaluation Review*. 32 (2): 187-215.

## 13. Author Biography

**Richard Berk, PhD**, formerly a Distinguished Professor of Statistics at UCLA, is a Professor of Criminology and Statistics at the University of Pennsylvania. He works on a wide variety of issues in criminology: inmate classification and placement systems, law enforcement strategies for reducing domestic violence, the role of race in capital punishment, detecting violations of environmental regulations, claims that the death penalty serves as a general deterrent, and forecasting short-term changes in urban crime patterns. He is equally active on a range of methodological concerns: causal inference, statistical learning, and methods for evaluating social programs. Professor Berk is an elected fellow of the American Association for the Advancement of Science, The American Statistical Association, and the Academy of Experimental Criminology. He has published 13 books and over 150 book chapters and articles. His most recent book is the controversial *Regression Analysis: A Constructive Critique* (Sage Publications, 2004). He currently finished up another book, *Statistical Learning from a Regression Perspective*, published in the Springer Series in Statistics, 2008.