Summary: Causality in Nonexperiments

How well do nonexperimental designs allow us to meet the criteria for identifying a nomothetic cause that were identified in Chapter 6?

Association between the hypothesized independent and dependent variables. Nonexperiments can provide clear evidence of association between the independent and dependent variables.

Time order of effects of one variable on the others. Cross-sectional designs (one-shot case studies) can establish time order with some confidence, but only in the limited cases identified in Chapter 6. Longitudinal designs, even when nonexperimental, do allow identification of time order.

Nonspurious relationships between variables. This is the Achilles heel of nonexperimental designs. As explained in Chapter 6, statistical controls can increase our confidence that extraneous variables do not confound the relationship between the independent and dependent variables, but it is unlikely that we will be able to control for all potential confounders.

Mechanism that creates the causal effect. Nonexperimental designs have no particular advantages or disadvantages for establishing causal mechanisms, although qualitative research designs facilitate investigations about causal process.

Context in which change occurs. Because they make it easy to survey large numbers of widely dispersed persons or organizations, one-shot case studies facilitate investigation of contextual effects.

VALIDITY IN EXPERIMENTS

Like any research design, experimental designs must be evaluated in terms of their ability to yield valid conclusions. True experiments are particularly well suited to producing valid conclusions about causality (internal validity), but they are likely to fare less well in achieving generalizability. Quasi-experiments may provide more generalizable results than true experiments but are more prone to problems of internal validity (although some quasi-experimental designs allow the researcher to rule out almost as many potential sources of internal invalidity as does a true experiment). Measurement validity is also a central concern, but experimental design does not in itself offer any special tools or particular advantages or disadvantages in measurement. In this section you will learn more about the ways in which experiments help (or don’t help) to resolve potential problems of internal validity and generalizability.

Causal (Internal) Validity

An experiment’s ability to yield valid conclusions about causal effects is determined by the comparability of its experimental and comparison groups. First, of course, a comparison group must be created. Second, this comparison group must be so similar to the experimental group...
or groups that it shows what the experimental group would be like if it had not received the experimental treatment—if the independent variable had not varied.

There are four basic sources of noncomparability (other than the treatment) between a comparison group and an experimental group. They produce four of the five sources of internal invalidity:

- **Selection bias.** When characteristics of the experimental and comparison group subjects differ.
- **Endogenous change.** When the subjects develop or change during the experiment as part of an ongoing process independent of the experimental treatment.
- **History effects.** When something occurs during the experiment, other than the treatment, which influences outcome scores.
- **Contamination.** When either the experimental group or the comparison group is aware of the other group and is influenced in the posttest as a result (Mohr, 1992).

The fifth source of internal invalidity can be termed **treatment misidentification:** Variation in the independent variable (the treatment) is associated with variation in the observed outcome, but the change occurs through a process that the researcher has not identified.

**Selection Bias**

You may already realize that the composition of the experimental and comparison groups in a true experiment is unlikely to be affected by **selection bias.** Randomization equates the groups’ characteristics, though with some possibility for error due to chance. The likelihood of difference due to chance can be identified with appropriate statistics.

But in field experiments, what has been planned as a random assignment process may deteriorate when it is delegated to front-line program staff. This problem occurred in the Sherman and Berk (1984) domestic violence experiment in Minneapolis. Some police officers sometimes violated the random assignment plan when they thought the circumstances warranted arresting a suspect who had been randomly assigned to receive just a warning. In several of the follow-up studies, the researchers maintained closer control over the assignment process so that randomization could be maintained.

Even when random assignment works as planned, the groups can become different over time because of **differential attrition,** or what can be thought of as “deselection.” That is, the groups become different because subjects are more likely to drop out of one of the groups for various reasons. This is not a likely problem in a laboratory experiment that occurs in one session, but some laboratory experiments occur over time, making differential attrition a problem. Subjects who experience the experimental condition may become more motivated than comparison subjects to continue in the experiment.

When subjects are not assigned randomly to treatment and comparison groups, as in nonequivalent control group designs, the threat of selection bias is very great. Even if the researcher selects a comparison group that matches the treatment group on important variables, there is no guarantee that the groups were similar initially in terms of the dependent variable or in terms of some other characteristic that ultimately influences posttest scores. However, a pretest helps the researchers to determine and control for selection bias. Because most variables that might influence outcome scores will also have influenced scores on the
pretest, statistically controlling for the pretest scores also serves to control for many of the unmeasured variables that might have influenced the posttest scores.

**Endogenous Change**

The type of problem subsumed under the label **endogenous change** occurs when natural developments in the subjects, independent of the experimental treatment itself, account for some or all of the observed change between pretest and posttest. Endogenous change includes these three specific threats to internal validity:

- **Testing.** Taking the pretest can in itself influence posttest scores. Subjects may learn something or be sensitized to an issue by the pretest and as a result, respond differently the next time they are asked the same questions, on the posttest.
- **Maturation.** Changes in outcome scores during experiments that involve a lengthy treatment period may be due to maturation. Subjects may age, gain experience, or grow in knowledge all as part of a natural maturational experience and thus respond differently on the posttest than on the pretest.
- **Regression.** People experience cyclical or episodic changes that result in different posttest scores, a phenomenon known as a **regression effect**. Subjects who are chosen for a study because they received very low scores on a test may show improvement in the posttest, on average, simply because some of the low scorers were having a bad day. On the other hand, individuals selected for an experiment because they are suffering from tooth decay will not show improvement in the posttest because a decaying tooth is not likely to improve in the natural course of things. It is hard in many cases to know whether a phenomenon is subject to naturally occurring fluctuations, so the possibility of regression effects should be considered whenever subjects are selected because of their extremely high or low values on the outcome variable (Mohr, 1992:56, 71–79).

Testing, maturation, and regression effects are generally not a problem in true experiments, because they would affect the experimental group and the comparison group equally. However, these effects could explain any change over time in most before-and-after designs, because they do not have a comparison group. Repeated measures panel studies and time series designs are better in this regard because they allow the researcher to trace the pattern of change or stability in the dependent variable up to and after the treatment. Ongoing effects of maturation and regression can thus be identified and taken into account.

**External Events**

History, or **external events** during the experiment (things that happen outside the experiment), could change subjects’ outcome scores. Examples are newsworthy events that have to do with the focus of an experiment and major disasters to which subjects are exposed. This problem is often referred to as a **history effect**—history during the experiment, that is.

Causal conclusions can be invalid in some true and quasi-experiments because of the influence of external events. For example, in an experiment in which subjects go to a special location for the treatment, something in that location unrelated to the treatment could influence these subjects. Experimental and comparison group subjects in the Price et al. (1992) study
of job search services differed in whether they attended the special seminars, so external events could have happened to subjects in the experimental group that might not have happened to those in the control group. Perhaps program participants witnessed a robbery outside of the seminar building one day, and their orientations changed as a result. External events are a major concern in evaluation studies that compare programs in different cities or states (Hunt, 1985:276–277).

**Contamination**

Contamination occurs in an experiment when the comparison group is in some way affected by, or affects, the treatment group. This problem basically arises from failure to control adequately the conditions of the experiment. When comparison group members are aware that they are being denied some advantage, they may as a result increase their efforts to compensate, creating a problem termed compensation rivalry, or the John Henry effect (Cook & Campbell, 1979:55). On the other hand, comparison group members may become demoralized if they feel that they have been left out of some valuable treatment and perform worse than expected as a result. Both compensation rivalry and demoralization thus distort the impact of the experimental treatment.

Contamination is not ruled out by the basic features of experimental and quasi-experimental designs, but careful inspection of the research design can determine how much it is likely to be a problem in a particular experiment. If the experiment is conducted in a laboratory, if members of the experimental group and the comparison group have no contact while the study is in progress, and if the treatment is relatively brief, contamination is not likely to be a problem. To the degree that these conditions are not met, the likelihood of contamination will increase.

Contamination was a potential problem in the field-based unemployment training study by Price et al. (1992), because all participants used the same unemployment offices.

**Treatment Misidentification**

Treatment misidentification occurs when some process that the researcher is not aware of is responsible for the apparent effect of treatment. The subjects experience something other than, or in addition to, the treatment the researchers believe they have experienced. Treatment misidentification has at least three sources:

- Expectancies of experimental staff: Change among experimental subjects may be due to the positive expectations of the experimental staff who are delivering the treatment rather than due to the treatment itself. Even well-trained staff may convey their enthusiasm for an experimental program to the subjects in subtle ways. This is a special concern in evaluation research, when program staff and researchers may be biased in favor of the program for which they work and eager to believe that their work is helping clients. Such positive staff expectations thus create a self-fulfilling prophecy. However, in experiments on the effects of treatments like medical drugs, double-blind procedures can be used: Staff delivering the treatments do not know which subjects are getting the treatment and which are receiving a placebo, something that looks like the treatment but has no effect.
Placebo effect. Treatment misidentification may occur when subjects receive a treatment that they consider likely to be beneficial and improve because of that expectation rather than because of the treatment itself. In medical research, where the placebo is often a chemically inert substance that looks like the experimental drug but actually has no effect, some research has indicated that the placebo effect itself produces positive health effects in many patients suffering from relatively mild medical problems (Goleman, 1993a:C3). It is not clear that these improvements are really any greater than what the patients would have experienced without the placebo [Hrobjartsson & Gotzsche, 2001]). In any case, it is possible for placebo effects to occur in social science research, so, when possible, experimental researchers can reduce this threat to internal validity by treating the comparison group with something that seems similar to what the experimental group receives.

Hawthorne effect. Members of the treatment group may change in terms of the dependent variable because their participation in the study makes them feel special. This problem could occur when treatment group members compare their situation to that of members of the control group, who are not receiving the treatment, in which case it would be a type of contamination effect. But experimental group members could feel special simply because they are in the experiment. This is termed a Hawthorne effect, after a famous productivity experiment at the Hawthorne electric plant outside Chicago. No matter what conditions the researchers changed in order to improve or diminish productivity, the workers seemed to work harder simply because they were part of a special experiment. Hawthorne effects are also a concern in evaluation research, particularly when program clients know that the research findings may affect the chances for further program funding.

Generalizability

The need for generalizable findings can be thought of as the Achilles heel of true experimental design. The design components that are essential for a true experiment and that minimize the threats to causal validity make it more difficult to achieve sample generalizability (being able to apply the findings to some clearly defined larger population) and cross-population generalizability (generalizing across subgroups and to other populations and settings).

Sample Generalizability

Subjects who can be recruited for a laboratory experiment, randomly assigned to a group, and kept under carefully controlled conditions for the study’s duration are unlikely to be a representative sample of any large population of interest to social scientists. Can they be expected to react to the experimental treatment in the same way as members of the larger population? The generalizability of the treatment and of the setting for the experiment also must be considered (Cook & Campbell, 1979:73–74). The more artificial the experimental arrangements, the greater the problem (Campbell & Stanley, 1966:20–21).

A researcher can take steps both before and after an experiment to increase a study’s generalizability. Conducting a field experiment, like Sherman and Berk’s (1984) study of arrest in actual domestic violence incidents, is likely to yield more generalizable findings than are laboratory experiments, for which subjects must volunteer. In some field experiments,
participants can even be selected randomly from the population of interest, and thus the researchers can achieve results generalizable to that population. For example, some studies of the effects of income supports on the work behavior of poor persons have randomly sampled persons within particular states before randomly assigning them to experimental and comparison groups. When random selection is not feasible, the researchers may be able to increase generalizability by selecting several different experimental sites that offer marked contrasts on key variables (Cook & Campbell, 1979:76–77).

**External Validity**

Researchers are often interested in determining whether treatment effects identified in an experiment hold true for subgroups of subjects and across different populations, times, or settings. Of course, determining that a relationship between the treatment and the outcome variable holds true for certain subgroups does not establish that the relationship also holds true for these subgroups in the larger population, but it suggests that the relationship might be externally valid.

We have already seen examples of how the existence of treatment effects in particular subgroups of experimental subjects can help us to predict the cross-population generalizability of the findings. For example, Sherman and Berk's research (see Chapter 2) found that arrest did not deter subsequent domestic violence for unemployed individuals; arrest also failed to deter subsequent violence in communities with high levels of unemployment. Price et al. (1992) found that intensive job-search assistance reduced depression among individuals who were at high risk for it because of other psychosocial characteristics; however, the intervention did not influence the rate of depression among individuals at low risk for depression. This is an important interaction effect that limits the generalizability of the treatment, even if Price et al.'s sample was representative of the population of unemployed persons.

There is always an implicit tradeoff in experimental design between maximizing causal validity and generalizability. The more that assignment to treatments is randomized and all experimental conditions are controlled, the less likely it is that the research subjects and setting will be representative of the larger population. College students are easy to recruit and to assign to artificial but controlled manipulations, but both practical and ethical concerns preclude this approach with many groups and with respect to many treatments. However, although we need to be skeptical about the generalizability of the results of a single experimental test of a hypothesis, the body of findings accumulated from many experimental tests with different people in different settings can provide a very solid basis for generalization (Campbell & Russo, 1999:143).

**Interaction of Testing and Treatment**

A variant on the problem of external validity occurs when the experimental treatment has an effect only when particular conditions created by the experiment occur. One such problem occurs when the treatment has an effect only if subjects have had the pretest. The pretest sensitizes the subjects to some issue, so that when they are exposed to the treatment, they react in a way they would not have reacted if they had not taken the pretest. In other words, testing and treatment interact to produce the outcome. For example, answering questions in a
pretest about racial prejudice may sensitize subjects so that when they are exposed to the experimental treatment, seeing a film about prejudice, their attitudes are different from what they would have been. In this situation, the treatment truly had an effect, but it would not have had an effect if it were repeated without the sensitizing pretest. This possibility can be evaluated by using the Solomon Four-Group Design to compare groups with and without a pretest (see Exhibit 7.10). If testing and treatment do interact, the difference in outcome scores between the experimental and comparison groups will be different for subjects who took the pretest compared to those who did not.

As you can see, there is no single procedure that establishes the external validity of experimental results. Ultimately, we must base our evaluation of external validity on the success of replications taking place at different times and places and using different forms of the treatment.

**COMBINING METHODS**

Innovative methodologists can combine elements of two or more research designs to overcome limitations of each. Qualitative methods can help to understand experimental effects, and experimental techniques can be added to survey research projects to allow stronger hypothesis tests.

**Process Analysis**

*Process analysis* is a technique for understanding how an experiment affects the dependent variable. It also can be used to avoid mistaken identification of the experimental treatment as the causal influence when in fact some other aspect of the experiment is responsible (Hunt, 1985:272–274). Periodic measures are taken throughout an experiment to assess