Chapter twenty one

Special Topics: Randomized Experiments

What does a randomized experiment look like?

What are the advantages of randomized experiments?

How do randomized experiments maximize internal validity?

What is a block randomized trial?

How does block randomization help increase statistical power?

How can covariates be used to help increase statistical power?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_21,

Many of the descriptive and inferential statistical approaches we have examined in this text are appropriate for analyzing randomized experiments. For example, in Chapter 11 we used the two-sample test of proportions to analyze data from the Maricopa County Drug Testing Experiment. In this chapter we want to focus more specifically on randomized studies. We begin the chapter by showing why randomized experiments provide a very strong ability to make causal inferences without concern for confounding. Indeed, many scholars have taken the position that only randomized experiments can provide valid conclusions regarding the impacts of treatments and programs.[1] Joan McCord argued, for example, that crime and justice evaluations should employ random assignment "whenever possible."[2]

Once we have established why randomized experiments provide distinct advantages, we will focus on some specific approaches for strengthening analysis of

---

[1]See Robert Boruch, Brooke Snyder, and Dorothy DeMoya, "The Importance of Randomized Field Trials," *Crime & Delinquency* 46 (2000):156–180; Donald Campbell and Robert F. Boruch, "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects," In Carl A. Bennett and Arthur A. Lumsdaine (eds.), *Evaluation and Experiment*: *Some Critical Issues in Assessing Social Programs*, (New York: Academic Press, 1975: 195–296); Thomas Cook and Donald Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, (Boston, MA: Houghton Mifflin Harcourt, 1979); Lynette Feder, Annette Jolin, and William Feyerherm, "Lessons from Two Randomized Experiments in Criminal Justice Settings," *Crime & Delinquency* 46 (2000): 380–400; Brian R. Flay and J. Allen Best, "Overcoming Design Problems in Evaluating Health Behavior Programs," *Evaluation and the Health Professions*, 5 (1982): 43–69; David Weisburd, "Randomized Experiments in Criminal Justice Policy: Prospects and Problems," *Crime & Delinquency*, 46 (2000): 181–193; David Weisburd, Cynthia Lum, and Anthony Petrosino, "Does Study Design Affect Research Outcomes in Criminal Justice?" *The Annals of the American Academy of Political and Social Sciences* 578 (2001): 50–70; Leland Wilkinson and Task Force on Statistical Inference, APA Board of Scientific Affairs, "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist* 54 (1999): 594–604.

[2]See Joan McCord, "Cures that Harm: Unanticipated Outcomes of Crime Prevention Programs," *Annals of the American Academy of Political and Social Science* 587 (2003): 16–30, p. 29.

randomized studies. While the basic tools we have already described for analyzing data are also appropriate for experimental data, there are specific problems that researchers might encounter in experimental research. One of these derives from the fact that experiments are "carried out" by the researcher and practitioner in the field. Unlike "observational" research studies that observe programs and draw data from them, in a randomized experiment the researcher makes the scene by randomly allocating subjects or places to treatment and control conditions. This means that it is often difficult to gain enough cases for a statistically powerful research design (see Chapter 23). In this chapter we examine two methods for increasing the statistical power of experimental research programs. We also discuss a related problem that develops from the focus of experimental research on specific research problems. Some scholars have criticized randomized experiments because they generally do not take into account "interactions" between the variable or the treatment of interest and other relevant variables.[3] For example, in assessing a rehabilitation program in prison we may be interested not only in the general impacts of the program but also the differential impacts across men and women, or older or younger offenders. We will discuss how a researcher can do this in the context of an experimental research model.

The Structure of a Randomized Experiment[4]

The general structure of experiments in criminology is usually similar in design regardless of the area or the question of interest. Generally, experiments in criminology start with an eligibility pool, randomization, group allocation, and posttest measures relevant to the dependent variable of interest (Figure 21.1).

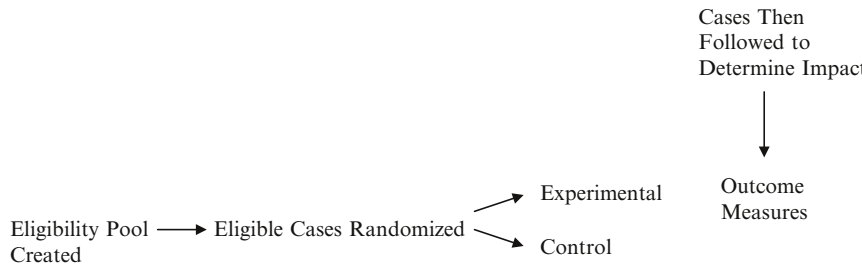
The eligibility pool is made up of those participants or units that are eligible for the experiment. Units of analysis can be individuals or aggregated groups or other entities that often are found in clusters. For example, in an experiment that evaluates the impact of increased foot patrol on crime rates, the eligibility pool may be individual officers who "walk the beat" in a specific area, and who have a specific number of years of experience. Or the unit of analysis could be the geographic area or the "beat" that will be assigned to different conditions. The eligibility pool thus comprises those patrol officers (or patrol "beats") that meet the criteria for inclusion in the study.

Next, researchers randomly assign members from this pool of eligible participants or units to the study conditions—often a treatment group and a control or a comparison group. Historically, randomization was carried out through a simple coin flip. There are now many ways to randomize subjects, but most often researchers rely on computerized statistical software to carry out randomization. Some researchers may simply use the rule of odds and evens—that is, assigning every other case to one particular group. This is often referred to as alternation and is

[3] James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experimentation," *Journal of Economic Perspectives*, 9 (1995): 85–110.

[4] For this section and the section on internal validity we draw heavily from David Weisburd, Anthony Petrosino, Trevor Fronius (2013). Randomized Experiments in Criminology and Criminal Justice. In David Weisburd and Gerben Bruinsma (Eds.). *Encyclopedia of Criminology and Criminal Justice.* New York: Springer Verlaag.

Figure 21.1 *Diagram of the Typical Criminological Experiment*



considered quasi-random assignment as the assignment of the numbers used is not actually random. The most critical factor in randomization, however, is that each case has the same probability or likelihood of being selected for the control group as the experimental group and that the assignment is based purely on chance.

In the usual criminological experiment, eligible cases are randomly assigned to one of the two groups—treatment or control. Experiments in criminology may have more than two groups. But typically, an experiment comprises a group that receives the treatment or the intervention and a control or a comparison group that does not. It is also quite common in a criminological experiment for the control group to actually receive something rather than nothing. For example, in a foot patrol experiment, the control group may receive treatment as usual or the same number of foot patrol officers as typically employed.

Experimental designs can include any number of outcomes. If the randomization was implemented with fidelity, it should produce two equivalent groups on the pretest or baseline measures related to the outcome of interest. The researcher then conducts analyses to determine if the intervention had any impact on the posttest or follow-up measures of the outcomes of interest.

The Main Advantage of Experiments: Isolating Causal Effects

In identifying whether a variable has a causal influence, or evaluating the outcomes of treatments or programs, the key issue for researchers is getting an unbiased estimate of the treatment or the intervention effect. Without that any other considerations, such as the ability to generalize results, are superfluous. For example, suppose an evaluator was asked to assess whether an intervention for drug involved offenders provided an effective deterrent to future offending. In the study employed, the treatment group was found to be half as likely to recidivate as the control condition not receiving the intervention. This would ordinarily lead the evaluator to report that the intervention was a success. But what if it was difficult to "believe" the result that was gained in the study? Suppose that the design of the study did not allow the evaluator to assume that the observed effect was actually the result of the intervention. In this scenario, the evaluator could not be sure that

it was the intervention that caused the change or something else that was common to the treatment group but not to the control group. In such a situation, it does not do much good to argue about whether the results can be generalized to a specific population of interest. The results themselves are not believable.

The main problem researchers face in producing believable results is that treatments are often confounded with other factors. For example, suppose that the reason for the outcome observed above was that the evaluator had not taken into account the fact that the treated drug offenders were volunteers. Volunteers in turn are more likely to be highly motivated to succeed in such programs than individuals who are not volunteers.[5] This is often termed "creaming" in the identification of the subjects in the treatment condition. The reason why the intervention group had lower recidivism rates in this case could easily be explained by the fact that they were on average more motivated to be rehabilitated than drug offenders in the control condition.

All research studies that seek to establish causation between a specific variable or treatment and an outcome must deal with this problem of confounding, and it stands as the major barrier to drawing believable conclusions in criminological studies. Non-experimental methods, such as regression techniques using observational data that we reviewed in Chapters 15–20, and quasi-experiments using approaches such as matching of subjects rely on a similar logic to solve the problem of confounding. The logic is easily stated: if we know what the factors are that confound treatment (or the variable of interest) we can take them into account. In other words, non-experimental methods, as we noted in Chapter 16, rely on a "knowledge solution" to the problem of confounding.

But how does knowledge solve the problem of confounding? Let us take the example of regression analyses as described in Chapter 16 using observational data examining the question of the effect of a drug intervention program on recidivism. Figure 21.2a shows the effect of the intervention on recidivism using a standardized regression coefficient approach (see pages 492–493). Here we have the simple relationship between treatment and recidivism with a coefficient value of −.25, a result suggesting that the intervention decreases recidivism. However, when we include in our analysis the "confounding" factor—level of motivation—the relationship between treatment and recidivism changes (see Figure 21.2b). Taking into account the effect of level of motivation, the effect of the intervention becomes 0 in this illustration. The observed effect was not due to the intervention but rather due to the confounding of the intervention with motivation of offenders.

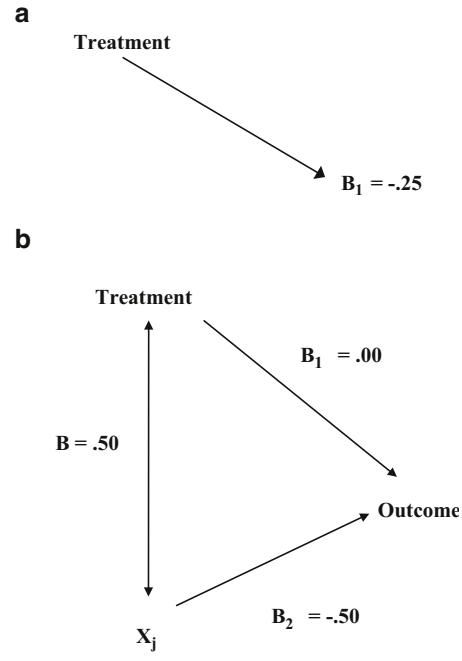
Notice that two extra coefficients are included in Figure 21.2b. The first represents the relationship between treatment (the variable of interest) and motivation (the confounding variable). This standardized coefficient is .50 and represents the extent to which treatment and motivation are related or confounded. The second

---

[5]See George De Leon, Gerald Melnick, George Thomas David Kressel, and Harry K. Wexler, "Motivation for Treatment in a Prison-Based Therapeutic Community," *American Journal of Drug and Alcohol Abuse*, 26 (2000): 33–46.; Faye Taxman, *Reducing Recidivism through a Seamless System of Care: Components of Effective Treatment, Supervision, and Transition Services in the Community*, (Washington, DC: Office of National Drug Control Policy, 1998).; Robert Rosenthal, "The Volunteer Subject," *Human Relations* 18 (1965): 389–406.

Figure 21.2

Example of the Bias in the Estimate of a Treatment Effect Caused by the Exclusion of an Unknown or an Unmeasured Factor (X_j). (a) Estimate of B_1 in the Case Where the Factor (X_j) Is Unmeasured and Excluded from the Model. Estimate of B_1 is −.25. (b) Estimate of B_1 in the Case Where the Factor (X_j) Is Included in the Model. Estimate of B_1 is .00.



additional coefficient (−.50) represents the relationship between motivation and recidivism. Together these two relationships detail the extent of confounding that is clouding our ability to estimate the treatment or the program effect. Confounding in this context takes into account the extent to which the confounding factor is related to the outcome of interest, and the degree to which it is related to or confounded with the treatment variable. And indeed, we can estimate the overall confounding by simply multiplying them. This gives us a value of −.25, the value of the simple relationship observed in Figure 21.2a. This can be defined as the degree of bias. In this case, the degree of bias is equal to the observed effect. Had we not taken motivation into account we would have erroneously concluded that the treatment leads to lower rates of recidivism, when in fact it is the motivation of offenders (represented by **Xj**) which is responsible for this result.

The way in which multivariate regression approaches allow us to control for confounding is illustrated in Equation 21.1 and described in Chapter 16. Here we show the computation of the regression coefficient b for a treatment variable (Tr) controlling for a confounding factor (CC, in this case motivation):

$$b_{Tr} = \left(\frac{r_{Y_i Tr} - \left(r_{Y_i CC} r_{TrCC} \right)}{1 - r^2_{TrCC}} \right) \left(\frac{S_Y}{S_{Tr}} \right) \qquad \textbf{Equation 21.1}$$

The key part of the equation for our interest is the numerator in the first part of the equation: $r_{Y_1Tr} - (r_{Y_1CC} r_{TrCC})$. Note that it includes the simple correlation between the treatment variable and the outcome measure (r_{Y_1Tr}). Subtracted from that is the product of the correlation between the outcome measure and the confounding variable (r_{Y_1CC}) and the treatment and confounding variable (r_{TrCC})—the two components of confounding we have just described. In this context we can statistically "un-confound" our estimate of the treatment if we have knowledge of the confounding factor. In our example the computation would be $-.25 - (.50 * -.50)$ or $-.25 + .25$, or 0, suggesting a null effect for treatment.

This solution is also key to the myriad of approaches that have been developed for other types of non-experimental approaches. All of them rely on knowledge about confounding. For example, matching of subjects on known characteristics, or the more sophisticated propensity score matching approach,[6] begins with the basic assumption that we have enough knowledge to create equivalence of units in the treatment and control conditions. Because the subjects in the matched groups are assumed to be alike, we make an assumption that confounders are not influencing our observations of a treatment effect. Note that in this case we are trying to statistically control for such confounding factors, by making the treatment and control groups alike on these factors. In such a case, the correlation between treatment and confounding variables is assumed to be 0. When it is, as illustrated in Equation 21.2 for the bivariate regression coefficient (see Chapter 15), the effect of the treatment breaks down to the simple correlation between treatment and outcome:

$$b_{Tr} = \left(r_{Y_1Tr}\right)\left(\frac{S_Y}{S_{Tr}}\right) \qquad \textbf{Equation 21.2}$$

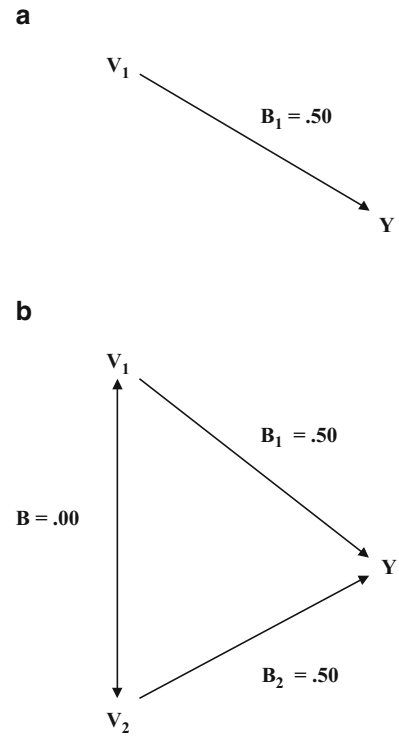
The problem with these methods is that if we want to get an unbiased estimate of treatment we would in theory have to identify all "confounding causes." Using the regression approach, which in some sense provides the most transparent form of non-experimental methods, we would need to identify all confounding factors that also have meaningful impacts on the outcome measure and include them in the regression. This would mean both that we would have to have knowledge about all such confounding factors and that we would be able to measure them in a research study.

Randomized experiments start with a different logic. If we cannot control out for confounding, we can make it irrelevant for the problem at hand. This is done through the process of randomizing treatment. If treatment is randomized then there is no reason to suspect systematic biases. This can be illustrated by returning to the simple path diagrams we used earlier. In Figure 21.3a we show the simple relationship between a treatment and outcome. In Figure 21.3b we include a potential confounding variable. Note that the confounding factor has a strong

[6] See Paul R. Rosenbaum and Dennis R. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1983): 41–55.; Paul R. Rosenbaum and Dennis R. Rubin, "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79 (1984): 516–524.

Figure 21.3

Example of the Lack of Confounding in the Treatment Effect When the Treatment (V_1) and Potential Confounder (V_2) Have No Relationship. (a) The Model Excluding a Potential Confounder, V_2. (b) The Model Including a Potential Confounder (V_2) but No Relationship Between the Treatment (V_1) and the Confounder Because of Randomization. If V_2 is Excluded, the Bias = $B * B_2 = .00 * .50 = .00$



standardized relationship with the outcome variable (B = .50). However, using the theory of randomization we can assume that there is no systematic relationship between the treatment and the confounder. This is the case because treatment has been randomly allocated. In theory it is not going to be related systematically to other factors such as gender, race, age, and attitudes.

What this means is that the relationship between any potential confounder and the treatment can be assumed to be 0. By chance, fluctuations will occur, and there will be systematic relationships observed, but these can in this case be assumed to balance out in the long run. Or at least there is no reason to assume that they will not. And if the relationship between the confounder and the treatment is 0, then when we multiply this by the large relationship between the confounder and the outcome we will also gain 0. The effect of treatment is not confounded. This is also illustrated in Equation 21.2 presented earlier, though in this case the assumption that treatment and the confounder have a 0 correlation is more believable.

Internal Validity

Our discussion so far is often subsumed under the heading of "internal validity" in methodological texts in criminology. A research design in which the impact of the intervention can be clearly distinguished from other observed factors is known as having high internal validity. If there are confounding factors involved in the impact of the intervention, then the evaluation design is considered to have low internal validity. Shadish, Cook, and Campbell, among others, have identified the most common threats to internal validity[7]:

1. *Selection*: The preexisting differences between treatment and control subject or units.
2. *History*: An external event occurring at the same time of the study that may influence impact.
3. *Maturation*: Changes in subjects or units between measurements of the dependent variable. These changes may be of natural evolution (e.g., aging) or due to time-specific incidences (e.g., fatigue, illness).
4. *Testing*: Measurement at pretest impacts measurement at posttest.
5. *Instrumentation*: Changes to the instrument or the method of measurement in posttest measures.
6. *Regression to the mean*: Natural trends may cause extreme subjects or units who score extremely high or low during the pretest to score closer to the mean at posttest.
7. *Differential attrition*: The differential loss of subjects or units from the treatment group compared to the control group.
8. *Causal order*: The certainty that the intervention did in fact precede the outcome of interest.[8]

To further illustrate the importance of internal validity, let us suppose a researcher is interested in evaluating the impact of a youth court on juvenile recidivism. The internal validity is considered high if, at the end of the evaluation, the researcher can show that the change in juvenile recidivism among the intervention group is due only to the intervention (i.e., youth court) and no other confounding factors were at play. The researcher must show through either research design or analytical procedures that all confounding factors are accounted for in the measurement of outcomes. If the researcher is unable to account for other factors such as seriousness of first offense or the maturation of the study population, he or she must note that the observed effects may be due to other factors. If threats to validity (or potential confounding factors) are not accounted for, the internal validity of the study would be considered low.

---

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

[8]David P. Farrington and Brandon C. Welsh, "Randomized Experiments in Criminal Justice: What Have we Learned in the Past Two Decades?" *Journal of Experimental Criminology* 1 (2005): 9–38.

Generally speaking, a randomized experiment has the highest possible internal validity, because as we illustrated above, this approach allows the researcher to assume that other confounding causes of the outcome of interest, known and unknown, are not systematically influencing the study results. High internal validity in randomized experiments is gained through the process of randomly allocating the treatment or the intervention to the experimental and control or comparison groups. Through random assignment, the researcher is not just randomizing the treatment. He or she is randomizing all other factors that may influence the outcome of the treatment. Thus, there is no systematic bias that increases the odds of one unit's assignment to the treatment group and another unit's assignment to the control or the comparison group. This is not to imply that the groups are the same on every characteristic—it is very possible that differences may occur; however, these differences can be assumed to be randomly distributed and are accounted for in the probability distributions that underlie statistical tests of significance. Regardless, neither the treatment group nor the control group should have an advantage over the other on the basis of known or unknown variables. Thus, randomized experiments are the only design that allows the researcher to assume statistically unbiased effects.[9]

The goal of most randomized experiments in criminology and criminal justice, as in other social science fields, is to disentangle the impact of the treatment or the intervention from the impact of other factors on the outcomes that are to be tested. A randomized experiment allows the researcher to attribute differences between the groups from pretest to posttest to the treatments or the interventions that are applied. At the conclusion of the study the researcher is able to assert, with confidence, that the differences are likely a result of the treatment and not due to other confounding factors. It is more difficult for non-randomized studies, even a high-quality quasi-experimental design, to make this assertion. This advantage is underscored by Farrington[10]:

> The unique advantage of randomized experiments over other methods is high internal validity. There are many threats to internal validity that are eliminated in randomized experiments but are serious in non-experimental research. In particular, selection effects, owing to differences between the kinds of persons in one condition and those in another, are eliminated.

Sample Size, Equivalence, and Statistical Power[11]

Despite the distinct advantages of randomized studies, it is often difficult to gain a large number of cases in a randomized experiment. Sometimes this is true because it is difficult to identify a large number of subjects who can be made eligible for

[9] See Robert F. Boruch, *Randomized Experiments for Planning and Evaluation: A Practical Guide*, (Thousand Oaks, CA: Sage, 1997).

[10] See David P. Farrington "Randomized Experiments on Crime and Justice," In Michael Tonry (ed.), *Crime and Justice: A Review of Research*, 4 (1983): 257–308, p. 260.

[11] Our discussion in this section relies heavily on David Weisburd and Charlotte Gill, "Block Randomized Trials at Places: Rethinking the Limitations of Small N Experiments," *Journal of Quantitative Criminology* (2013).

randomization into treatment and control conditions. Sometimes this is the case because treatment conditions or data collection are expensive, and each new case will increase the cost of the study. These problems are particularly acute in place-based randomized trials since the number of places with a specific crime problem is generally limited.[12] Moreover, place-based trials ordinarily demand significant treatment resources per site, and accordingly it is expensive for agencies to "treat" a large number of sites at one time.[13]

Farrington[14] argues that small N studies are not likely to achieve realistic pretest balance across measured and unmeasured covariates.[15] This of course undermines the main advantage of experimental studies—that the control and treatment groups can be assumed to be equivalent and differ only in the receipt or the non-receipt of treatment.

While Farrington and colleagues have focused primarily on the problem of equivalence, a related criticism of the small N sizes of many randomized experiments can also be raised. If the sample sizes used for such studies are small, then their statistical power under traditional assumptions is also likely to be low. Statistical power (see Chapter 23) is a particularly important component of evaluation studies, because it assesses whether the study will provide a "fair test" of the interventions examined. Sample size is one of the key elements of statistical power, and experiments with very small samples are also likely to have low statistical power.

A design approach called "block randomization" provides a potential solution to the risk of unbalanced samples in small N studies as well as a valid method for increasing the statistical power of such studies. Block randomized experiments take

---

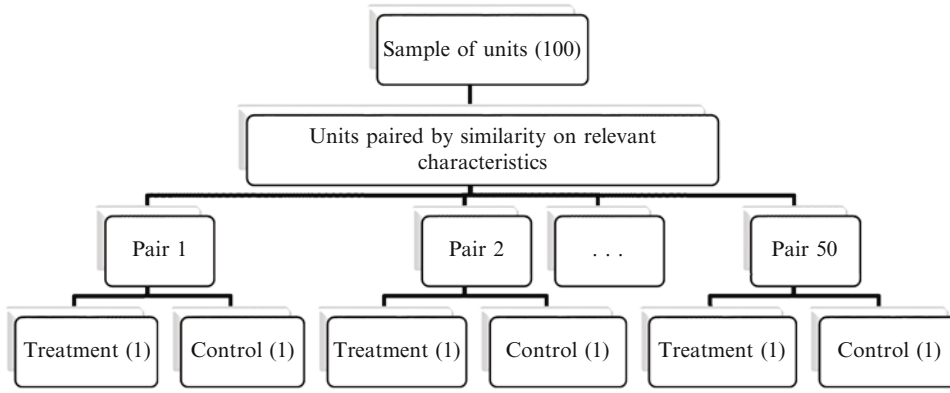
[12] See Anthony A. Braga, David L. Weisburd, Elin J. Waring, Lorraine Green Mazerolle, William Spelman, and Francis Gajewski, Problem-Oriented Policing in Violent Crime Places: A Randomized Controlled Experiment," *Criminology* 37 (1999): 541–580; Lawrence W. Sherman, Patrick R. Gartin, and Michael E. Buerger, "Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place," *Criminology* 27 (1989): 27–56.

[13] See Robert Boruch, Henry May, Herbert Turner, Julia Lavenberg, Anthony Petrosino, Dorothy De Moya, Jeremy Grimshaw, and Ellen Foley, "Estimating the Effects of Interventions that are Deployed in Many Places: Place-Randomized Trials," *American Behavioral Scientist* 47 (2004): 608–633; David Weisburd, "Hot Spots Policing Experiments and Criminal Justice Research: Lessons from the Field," *Annals of the American Academy of Political and Social Science* 599 (2005): 220–245.

[14] See note 8; See also David P. Farrington, Lloyd E. Ohlin, and James Q. Wilson, *Understanding and Controlling Crime: Toward a New Research Strategy*, (New York: Springer-Verlag, 1986); David P. Farrington and Maria M. Ttofi, "School-Based Programs to Reduce Bullying and Victimization," *Campbell Systematic Reviews*, 6(6), 2009; Darrick Jolliffe and David P. Farrington, *A Rapid Evidence Assessment of the Impact of Mentoring on Reoffending*, (London: Home Office Online Report, 2007).

[15] Farrington notes in this regard, "(t)o understand why randomization ensures closer equivalence with larger samples, imagine drawing samples of 10, 100, or 1,000 unbiased coins. With 10 coins, just over 10 percent of the samples would include 2 or less, or 8 or more, heads. With 100 coins, just over 10 percent of the samples would include 41 or less, or 59 or more, heads. With 1,000 coins, just over 10 percent of the samples would include 474 or less, or 526 or more, heads. It can be seen that, as the sample size increases, the proportion of heads in it fluctuates in a narrower and narrower band around the mean figure of 50 percent" (see note 8 p. 263n).

Figure 21.4 *Fully Blocked (Matched Pairs) Random Assignment*



advantage of prior knowledge about the distribution of units in an experimental study to maximize equivalence of treatment and control conditions.

We begin our discussion below by focusing on statistical theory, detailing how block randomized trials maximize equivalence of experimental studies and provide valid methods for increasing statistical power in small sample experiments. We then turn to an empirical illustration of these arguments, drawing from data used in the Jersey City Drug Market Analysis Experiment (Jersey City Experiment (JCE)).[16] Using data from the JCE and simulation methods, we illustrate the advantages of block randomization approaches over simple or "naïve" randomization in developing equivalent groups. We then illustrate the overall increase in statistical power provided by the block randomized statistical modeling approach.

Statistical Foundations for Block Randomization

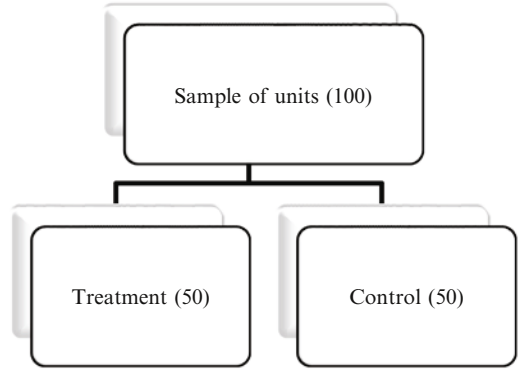
While criminological researchers are often unfamiliar with block randomized studies, the first randomized experiment in crime and justice, the Cambridge-Somerville Youth Study, used a complete or a fully blocked randomized design.[17] In that study, problem youths were paired on age, social background, biological somatotype, and temperament. The researchers used this paired or fully blocked design because the experimental treatment was lengthy and complex, so they sought to maximize the equivalence of the comparisons they could make. Their design is illustrated in Figure 21.4. In practice, the researchers matched the youths into pairs on these characteristics and then randomly allocated them within the pairs into treatment and control conditions. In the fully blocked randomized design, each subject in a pair has an equal probability of being assigned to treatment or control conditions, but randomization is restricted in that one subject from each pair must be assigned to treatment and one subject to control.

---

[16] See David Weisburd and Lorraine Green, "Policing Drug Hot Spots: The Jersey City Drug Market Analysis Experiment," *Justice Quarterly* 12 (1995): 711–735.

[17] See Edwin Powers and Helen Witmer, *An Experiment in the Prevention of Delinquency*, (New York: Columbia University Press, 1951).

Figure 21.5 *Naïve (Balanced) Random Assignment*



What advantage does this approach give over naïve or simple randomization? Naïve randomization (illustrated in Figure 21.5), which is the most common approach in crime and justice experiments, assigns the total sample under study to treatment or control conditions without restrictions. Every subject in this case has an equal probability of being assigned to either the treatment or the control condition. Naïve randomization relies on the assumption that there are no systematic reasons for the treatment and control subjects to differ (since every subject had an equal probability of assignment to each condition), a key *raison d'etre* for experimental studies in the first place. But it does not guarantee equivalence, simply that there is no reason for non-equivalence. When samples are large, this assumption is reasonable because large differences between the groups are unlikely by chance.

But why shouldn't we increase equivalence if we can, especially in small studies where chance differences between control and treatment groups might be large in the case of naïve randomization? Fully blocked randomized designs like the Cambridge-Somerville Youth Study assume that we have knowledge about the subjects or the units in an experiment that can help us create equivalence on factors that are related to the outcomes observed. Age and social background were considered key predictors of delinquency by the Cambridge-Somerville researchers, and their introduction as factors to match the youths in the study was seen as a direct way of making sure that the treatment and control conditions were similar on important influences of treatment success.

However, the benefit of equivalence gained through fully blocked randomization comes at a statistical price. For each limitation on randomization the study must "pay a fine" in terms of degrees of freedom. For example, in the Cambridge-Somerville Youth Study 650 boys were matched into pairs. In a naïve randomization design with 325 cases per group, the study would have had 648 degrees of freedom for statistical tests of significance ($N_1 + N_2 - 2$). Using the fully blocked design, the degrees of freedom of the tests declined to 324 ($N_{pairs} - 1$).

The loss of degrees of freedom is meaningful because it changes the distribution of the test statistic. For example, as illustrated by Equation 21.3 (dependent samples) and 21.4 (independent samples) below, in a *t*-test the estimated standard

deviations are divided by the degrees of freedom. This means that as the degrees of freedom of a test get smaller the *t*-value observed also gets smaller:

$$t = \frac{\bar{X}_d - \mu_d}{\sqrt{\frac{s_d^2}{df}}} \qquad \textbf{Equation 21.3}$$

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{df}}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}} \qquad \textbf{Equation 21.4}$$

In turn, the value of the statistic needed to achieve statistical significance will be larger as the degrees of freedom for a test get smaller (see the *t* distribution example in Table 21.1 below). This difference is not meaningful in the case of relatively large sample studies. For example, in the Cambridge-Somerville experiment, with 324 degrees of freedom, the critical value of the *t*-test (with standard criteria of $p < .05$ and a nondirectional test) is about 1.967, almost the same as the 1.960 in the *z* normal distribution without adjustment. But when the degrees of freedom are reduced to 100, the critical value for the *t*-test becomes 1.984 and at 50 degrees of freedom, 2.009.

The balance between loss of degrees of freedom and greater equivalence is weighted toward the goal of equivalence in disciplines where the causal processes underlying the impacts of treatment are well understood. This is the case because the benefits of the complete or the fully blocked randomized design are greatest when each loss of degrees of freedom is accompanied by a gain in the equivalence of the treatment and control conditions on factors that are related to treatment outcomes. If treatment outcomes are conditioned by such factors, then blocking will decrease the heterogeneity of outcomes in the study. Looking at Equation 21.3 and Equation 21.4 above, this would mean that the numerators of the standard errors are made smaller and accordingly the *t*-values observed are larger. This makes intuitive sense because if the groups are more similar in terms of what would have been expected absent treatment, then it should be easier to identify a treatment outcome. In statistical terms, there is likely to be less noise in identifying that outcome. In the case of a fully blocked design in which treatment outcomes

Table 21.1 Critical Values for the *t* Distribution (Two-Tailed, $\alpha = .05$)

DEGREES OF FREEDOM	CRITICAL VALUE
10	2.228
20	2.086
50	2.009
100	1.984
200	1.972
324	1.967
500	1.965
648	1.964

were not related to the blocking factors, the standard deviations would remain the same as in a naïve design, while there would be a substantial loss of degrees of freedom. This would mean that a large price was paid for the fully blocked design without a corresponding benefit.

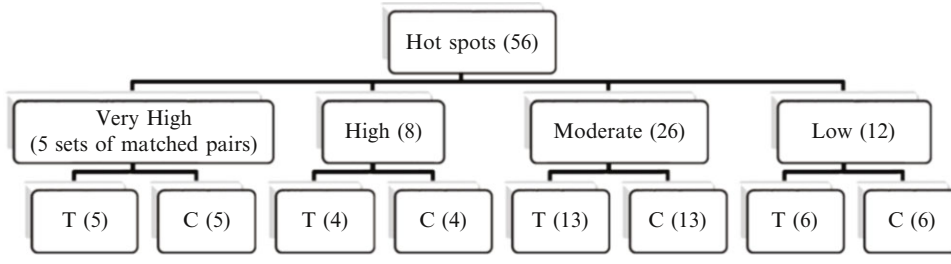
And here lies the primary argument against the use of fully blocked randomized designs in crime and justice. The level of knowledge of the causal processes underlying crime and justice research simply does not allow us to parse randomization with sufficient distinctions to allow us to gain a benefit from a fully blocked randomized design. This is one of the main reasons that matched pair designs are not common more generally in criminology, though we suspect that criminologists are often uninformed about the benefits of fully blocked randomized designs.

To say that criminologists do not have a full understanding of causal processes in place-based studies does not mean that they do not have sufficient knowledge to improve experimental designs using block randomization methods. A compromise approach between the fully blocked and naïve randomization approach is a partially blocked or termed here simply, block randomized design. A block randomized design makes no assumptions regarding the number of blocks or groups identified at the outset. Rather, the number of blocks is determined by the researcher's assessment of the ability of known data to group units in ways that maximize their similarities on key causal variables. Cases are placed within the specified randomization blocks and then randomized within those blocks (see Figure 21.6). The blocks do not have to be of equal size, but the number of cases in each block must be even to allow for equal randomization and balance within blocks.

An Example: Jersey City Experiment

At this juncture it is worthwhile to introduce the substantive example we will draw from for illustrating the benefits of a block randomized design for place-based studies, the Jersey City Drug Market Analysis Experiment. The JCE evaluated an innovative drug enforcement strategy involving police crackdowns along with citizen and local business engagement in controlling crime at drug markets. A total of 56 high drug activity hot spots were randomly assigned in equal numbers to receive either the experimental program or regular, unsystematic enforcement on an ad hoc basis. Most of the drug market hot spots included fewer than four street segments and intersections, though two places included more than ten street segments. Police emergency calls for service for a variety of crime and

Figure 21.6 *Partially Blocked Random Assignment (Jersey City Drug Market Analysis Experiment)*



disorder-related issues were measured for 7-month pre- and post-intervention periods. We focus below on three main outcome measures for disorder measured in the study: suspicious persons, public morals, and police assistance.

Knowing that there was considerable variation in criminal activity even across the sample of hot spots, the study authors were concerned at the outset that the prior level of crime would influence the effect of treatment. Given the small sample of drug hot spots that could be identified in Jersey City, the authors were also concerned that naïve randomization might lead to non-equivalent groups. At the same time, there was concern that each loss of degrees of freedom in the experiment would substantially impact the results, since the total N of available cases was only 56. The solution in the JCE was to examine the distribution of both emergency calls for service and arrests and then to identify natural cutting points.

In this way the researchers believed that they could gain greater equivalence between the groups without a large loss of degrees of freedom (28) that would have ensued if the fully blocked randomized design was adopted. The assumption here was that prior crime and disorder would have a general impact on the effects of treatment but would not be specific enough to distinguish sites in a way that would justify a complete randomized block design. The researchers identified eight statistical blocks for randomization. The ten highest activity hot spots were randomized in pairs because of large gaps between them; these five pairs represented the five "very-high-activity" statistical blocks. Of the rest of the sample of hot spots, 8 were grouped into a "high-activity" block, 26 hot spots were classified as a medium-activity block, and 12 were classified as a low-activity block.

The Benefits of Block Randomized Trials

One approach to examining the contribution of statistical blocking to equivalence in the Jersey City Drug Market Analysis Experiment is to compare the equivalence gained between the treatment and control conditions on key baseline (pretest) measures. However, the Jersey City study is only one specific draw of randomization. By definition any specific draw of a sample is going to be different from another draw. The statistical concern is whether on average, a draw using the block randomization procedure is likely to produce a more equivalent outcome than a draw using a simple randomization procedure. To examine this question we develop 10,000 simulations of both naïve randomization and block randomization using the Jersey City data.[18] We focus on baseline calls for service for the three key disorder outcomes in the study (suspicious persons, public morals, and assistance). Table 21.2 reports the baseline information from the original study, the simulation results for the blocking approach, and the simulation outcomes of a naïve

[18] Stata programs were developed to run a randomization sequence (blocked or naïve) on the JCE dataset and then run a *t*-test comparing the treatment and control group means at baseline on the three outcomes of interest. Stata's simulation function was then used to run each program 10,000 times and create a dataset containing the group means, *t*-values, *p*-values, an indicator showing whether or not the two groups were significantly different at baseline for each iteration, and the absolute average mean group difference across all iterations. We are grateful to David B. Wilson for developing the programs and simulation syntax.

Table 21.2 Calls for Service at Baseline in Block and Naïve Randomizations of JCE Data

	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
Original JCE block randomized data			
Treatment mean (SD)	17.00 (16.15)	9.32 (10.58)	43.86 (43.40)
Control mean (SD)	17.93 (21.16)	9.11 (13.36)	42.11 (43.05)
Absolute mean difference (SE)	.93 (5.03)	.21 (3.22)	1.75 (11.55)
10,000 simulations Block randomized data (N = 56)			
Average absolute mean difference across all samples (SD)	2.67 (1.95)	1.29 (0.94)	7.11 (5.05)
Number of samples with significant difference at baseline ($p \leq 0.10$)	93	4	218
10,000 simulations Naïve randomized data (N = 56)			
Average absolute mean difference across all samples (SD)	3.98 (2.89)	2.56 (1.86)	9.13 (6.71)
Number of samples with significant difference at baseline ($p \leq 0.10$)	955	989	966

randomization approach. In the case of the simulations, we report the number of samples that have significant differences at baseline and the overall absolute mean difference found in the 10,000 simulation samples.

Table 21.2 suggests the importance of the simulation approach. The specific draw in the Jersey City study produced unusually equivalent groups on the three baseline measures examined. The absolute mean difference for all three measures is substantially lower than the average absolute mean difference produced in our simulations. This does not mean that the Jersey City randomization was flawed but rather that the investigators by chance gained one of the more equivalent randomizations from the sampling distribution of all possible randomizations.

But despite the fact that the Jersey City randomization was a relatively "lucky draw," it is clear from Table 21.2 that the procedure used was likely to produce much more equivalent groups than a simple randomization procedure. In the 10,000 simulations of the JCE block randomization procedure only 93 simulations produced significantly different outcomes ($p < .10$) for treatment and control conditions at baseline for suspicious persons calls, 4 for public moral calls, and 218 for assistance calls. In contrast, using the simple randomization approach on the same 56 cases, 955 samples produced significant differences for suspicious person calls, 989 for public moral calls, and 966 for assistance calls.[19] These differences are of substantial magnitude and are also reflected in the average absolute mean difference across all of the simulation samples. For suspicious persons the mean differences were almost 50 % larger in the naïve randomization sample, for public morals about twice as large, and for assistance almost a third larger.

[19] Of course, this is about what we would have expected given a .10 significance threshold and a fair randomization procedure. But the important point is that the block randomization approach allows us to do better.

Statistical Power and Block Randomization

As we noted earlier, the benefits of a block randomized design are dependent on the assumption that the blocking factors are related to the study outcomes. This cannot be assessed ordinarily because knowledge about treatment outcomes is unknown until the experiment is complete. However, we are able to examine this assumption using post-experiment data from Jersey City. For all three outcomes of interest discussed here, the correlations between pre- and posttest outcomes were significant and had greater than a moderate size coefficient.[20] For suspicious persons the correlation was .44 ($p < .10$),[21] for public morals .52 ($p < .01$), and for assistance .63 ($p < .001$). These results suggest that Weisburd and Green's assumption that there would be a strong relationship between the blocking factors and the final study outcomes was correct.

In turn, the statistical model benefits of the identification of block variability can be observed directly in these data. A simple or a naïve randomized experiment presents a model for understanding outcomes where systematic variation is determined only by treatment. Accordingly, the model can be expressed in terms of sums of squares (*SS*; see Chapter 12) by Equation 21.5:

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{error} \qquad \textbf{Equation 21.5}$$

The total variability of the study in this case is broken down to the influence of the treatment (SS_{group}) and the overall variability in the data (SS_{error}), with the intercept ($SS_{intercept}$) completing the linear model. Remember that the statistical denominator for the statistical significance test is going to be the error term, meaning that as the error term gets smaller the significance of the test will get larger.

With the introduction of a blocking factor, an additional source of variability is taken into account in the model—SS_{block}—as illustrated in Equation 21.6:

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{block} + SS_{error} \qquad \textbf{Equation 21.6}$$

When the blocking factor (SS_{block}) is strongly related to the outcome (SS_{total}) we would expect the overall size of SS_{error} to decline. This is the case because the block randomized model limits any relationship between SS_{group} (i.e., the treatment component of the model) and SS_{block}, meaning that the two components of variability are constrained to be independent (because they are independent, the inclusion of the blocking factor in the model will not impact the size of the treatment effect). Accordingly, any SS_{block} effect will be drawn out of the error term for the model (SS_{total} is a fixed quantity irrespective of the model defined). Since SS_{error} is a key

[20] See Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., (Hillsdale, NJ: Lawrence Erlbaum, 1988).

[21] We calculated the correlation between the blocking factor and the three disorder outcome measures by running a GLM with only the blocking factor included. The correlation is based on taking the square root of the overall R2 of the model. We use a one-tailed test of significance following the assumption that the correlation between the blocking factor and the outcome is positive.

Table 21.3 Univariate Analysis of Variance for Treatment and Treatment–Block Effects (JCE)

UNIVARIATE ANALYSIS OF VARIANCE MODELS	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
Treatment-only model			
$SS_{intercept}$ *(df)*	480.286 (1)	21.875 (1)	1,512.161 (1)
SS_{group} *(df)*	516.071 (1)**	129.018 (1)	355.018 (1)
SS_{error} *(df)*	6,373.643 (54)	2,884.107 (54)	14,875.821 (54)
SS_{TOTAL} *(df)*	7,370.000 (56)	3,035.000 (56)	16,743.000 (56)
F (p) for group effect	4.372 (.041**)	2.416 (.126)	1.289 (.261)
Treatment and block model			
$SS_{intercept}$ *(df)*	510.998 (1)	145.484 (1)	3,250.491 (1)
SS_{group} *(df)*	516.071 (1)**	129.018 (1)*	355.018 (1)
SS_{block} *(df)*	1,352.076 (7)	803.346 (7)	6,087.060 (7)
SS_{error} *(df)*	5,021.566 (47)	2,080.761 (47)	8,788.761 (47)
SS_{TOTAL} *(df)*	7,370.000 (56)	3,035.000 (56)	16,743.000 (56)
F (p) for group effect	4.830 (.033**)	2.914 (.094*)	1.899 (.175)

Notes: *$p<.10$ **$p<.05$.

element of the denominator of the test statistic (the degrees of freedom being a second key element), its reduction without a proportionate reduction in SS_{group} (and a large decrease in degrees of freedom) will lead to a more significant outcome (i.e., a more powerful statistical outcome) than a naïve model.

As an illustration of these assumptions, we estimated univariate ANOVA models using just the treatment factor, and separately with the treatment and blocking factors as fixed effects (Table 21.3).[22] In order to simplify our example, we do not estimate the full model which could include a treatment by block interaction (see later).[23] Following our assumptions, SS_{total} and SS_{group} are the same in both models. The total variability in the model is constant irrespective of model specification, and the effect of treatment is not influenced because of the balanced randomization of cases within blocks. However, SS_{error} declines in the analyses that include blocking as a factor. For suspicious persons the decline is from 6,373.643 to 5,021.566, for public morals from 2,884.107 to 2,080.761, and for assistance from 14,875.821 to 8,788.761. Note as well that there is a corresponding decrease in the degrees of freedom of SS_{error} in the block randomized design (from 54 to 47 in all the models), reflecting the "price" of this approach. But, following our examination of the correlation between the blocking factor and the outcomes, the loss of statistical power generated by the reduction of degrees of freedom of the

[22] In this case the effects are fixed because we are assuming analysis of specific categories and do not assume that those categories are representative of the population of cases. For example, the analysis looks at the specific blocks of hot spots in the experiment; it does not assume that we have a representative sample of all possible "blocks" of hot spots. If the effects were random, the blocks observed would be seen as a representative sample of blocks of hot spots.

[23] Where the interaction between treatment and block is significant, Fleiss recommends including an interaction term in the model. When the blocking factor represents a substantively important variable, the introduction of a block by treatment interaction can also add knowledge about the differential effects of treatment across values of the blocking variable. See Joseph L. Fleiss, *The Design and Analysis of Clinical Experiments*, (New York: John Wiley and Sons, 1986); David Weisburd and Faye Taxman, "Developing a Multicenter Randomized Trial in Criminology: The Case of HIDTA," *Journal of Quantitative Criminology* 16 (200): 315–340.

error term is less than the gain from the inclusion of the blocking term. When we combine treatment and block effects in the model, all three comparisons show larger *F*-statistics. The observed *p*-value for suspicious persons declines from .041 to .033, for public morals from .126 to .094, and for assistance from .261 to .175.

Using Covariates to Increase Statistical Power in Experimental Studies

Another technique for increasing statistical power in experimental studies follows the statistical logic of block randomization but does not balance the blocking characteristics at the outset. It relies heavily on the logic of randomization that we have already described. As noted earlier, if the cases are randomized to treatment and control conditions then we can assume that there is no correlation between treatment and possible confounding factors (see Figure 21.2). That means that the inclusion of additional covariates in an analysis will not, in theory, affect the estimate of the treatment effect. Since that is the case, we should be able to include covariates without creating any bias in our assessment of the influence of the experimental variable.

However, we do gain a direct benefit in calculating the statistical significance of the test. Let us use again the approach of examining the sums of squares of our equation. Suppose we convert the JCE to a simple naïve randomization sequence. In this case our model includes treatment and error as the only variables (see Equation 21.5).

If we add covariates the error term for the model should decline, because as we noted earlier we have no reason to expect that the effect of treatment (i.e., group) will change (see Equation 21.7). As an example, let us add as covariates variables that should be related to the dependent variables: the pre-experiment calls for service for robbery and aggravated assault (collectively the baseline violent crime calls for service) and the baseline calls for service for each respective outcome measure. Thus, for each outcome, we include three covariates: robbery at baseline, aggravated assault at baseline, and the outcome at baseline (e.g., suspicious person calls at baseline for the suspicious person outcome):

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{covariates} + SS_{error} \qquad \textbf{Equation 21.7}$$

In Table 21.4 we show the results using the simple naïve design as well as the results we would gain for taking into account the three covariates for each outcome. As can be seen from the table, the statistical significance of the results including the covariates is considerably lower than when no covariates are included. For public morals, for example, the p-value for the group effect has dropped from a non-statistically significant .126 in the naïve example to a significant .041 when including the covariates. For all three outcomes, we have substantially lowered the SS_{error} by adding the covariates. Even though we paid a price in degrees of freedom for using three covariates, the benefit in terms of reducing the error and increasing the significance of our group findings outweighs the cost. We should

Table 21.4 Univariate Analysis of Variance for Treatment and Treatment–Covariate Effects (JCE)

UNIVARIATE ANALYSIS OF VARIANCE MODELS	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
Treatment-only model			
$SS_{intercept}$ (df)	480.286 (1)	21.875 (1)	1,512.161 (1)
SS_{group} (df)	516.071 (1)**	129.018 (1)	355.018 (1)
SS_{error} (df)	6,373.643 (54)	2,884.107 (54)	14,875.821 (54)
SS_{TOTAL} (df)	7,370.000 (56)	3,035.000 (56)	16,743.000 (56)
F (p) for group effect	4.372 (.041**)	2.416 (.126)	1.289 (.261)
Treatment and covariate model			
$SS_{intercept}$ (df)	188.025 (1)	34.122 (1)	84.815 (1)
SS_{group} (df)	559.666 (1)**	162.019 (1)**	638.146 (1)*
$SS_{covariate\ (pre-outcome)}$ (df)	678.443 (1)**	751.986 (1)**	305.731 (1)
$SS_{covariate\ (pre-robbery)}$ (df)	752.979 (1)**	160.373 (1)**	2403.850 (1)**
$SS_{covariate\ (pre-assault)}$ (df)	55.281 (1)	15.301 (1)	405.114 (1)
SS_{error} (df)	5,334.804 (51)	1,875.149 (51)	9596.472 (51)
SS_{TOTAL} (df)	7,370.000 (56)	3,035.000 (56)	16,743.000 (56)
F (p) for group effect	5.350 (.025**)	4.407 (.041**)	3.391 (.071*)

Notes: *p<.10 **p<.05.

be cautious in including these covariates, however. Note that we would expect the effect of treatment to remain similar between the simple model and the model with covariates in terms of the sums of squares explained. This is largely true for suspicious persons and public morals where the SS_{group} remains fairly similar in both sections of Table 21.4. For assistance, however, there is a large increase in the SS_{group} potentially suggesting that we may have introduced some level of bias into the model with our choice of covariates.

Importantly, if we had the data available, we could add many different covariates to the models. Those covariates could be related to the characteristics of the units of randomization, for example the social characteristics of the hot spots. In a study of individuals we might include gender or age or race to the analysis. In some sense, the larger the group of covariates that are related to the outcome measure that are included, the greater the benefit. This is because each additional variable included that is relevant to the prediction of the outcome will decrease the error term for the significance test. The rule does not apply if there is no relationship between the covariate and the outcome. In that case, the researcher will lose a degree of freedom (for every additional variable or parameter) in the analysis without an additional benefit in reduction of the error variance.

As is apparent, there is much to be gained by including covariates in an experimental analysis. However, as in other statistical procedures, covariates can be manipulated in ways that affect the validity of your results. Randomization allows us to assume that there is no relationship between the covariate and the treatment or the variable interest. But this does not mean that there is not in the sample of interest a spurious relationship that is observed. In any randomization there are likely to be some measures that by chance are related to the treatment. What randomization guarantees is that such bias will be random, and it is likely in the long run that whenever there is a spurious correlation it is likely to be balanced off with another correlation in the opposite direction. But what if the researcher

chooses specific variables that have relationships in the dataset with the treatment or the variable of interest? In this case, the error term will be reduced if these measures are correlated with the outcome, but so will the estimate of the treatment effect. Again, here as in regression analyses more generally, the researcher can "go fishing" until the result they are looking for is gained.

The dangers of influencing the validity of the treatment effect should lead to caution in using covariates to reduce error variance in the analysis of experimental studies. A general rule that will protect you from the danger of manipulation of results is for the researcher to define at the outset which covariates will be used in analyzing the outcomes. In this way, the researcher cannot manage results post facto on the basis of knowledge of sample characteristics. Clearly, one should not run a large number of regressions with different covariates included until a "good" result is gained. The process of selecting variables before the results of an experiment are known is in our view a good rule to follow. But more generally, if an experiment has sufficient statistical power, the researcher should use the simple analysis approach, in which covariates are not included. This is the only way to guarantee that the results are not being manipulated in a way that might lead to spurious findings.

Examining Interaction Terms in Experimental Research

Some scholars have criticized experimental studies because in their simplest form they do not allow us to examine contextual factors that might influence treatments or outcomes. For example, the average treatment effect observed in an experiment will tell you whether the treatment is effective or not for the entire sample. But what if we are interested in understanding whether the treatment effect differs for men and women or younger versus older subjects. Most experimental studies in criminology have not looked at these "interaction effects" between treatment and other factors. But this does not mean that interactions cannot be observed in a valid way in experimental studies.

The best way to examine interactions in an experimental study is to define the contextual factor at the outset. For example, in the JCE, the block randomization procedure was based on the level of crime observed in the baseline year. For illustration purposes we use four blocks in our analysis: very high, high, moderate, and low. One might ask whether the level of crime was related to the effectiveness of treatment. Importantly, because of the use of block randomization the cases are equally divided between the blocks, meaning that the design of the study ensures that there is no relationship between block and treatment. There are an equal number of very high, high, moderate, and low baseline crime hot spots in the experimental and control groups. We add to our equation an additional term beyond treatment or group and block—that of the interaction between treatment and block as illustrated in Equation 21.8:

$$SS_{total}=SS_{intercept}+SS_{group}+SS_{block}+SS_{group\times block}+SS_{error} \quad \textbf{Equation 21.8}$$

Table 21.5 Univariate Analysis of Variance for Treatment, Block, and Group by Block Effects (JCE)

UNIVARIATE ANALYSIS OF VARIANCE MODEL	SUSPICIOUS PERSONS	PUBLIC MORALS	ASSISTANCE
Treatment, block, and group × block model			
$SS_{intercept}$ *(df)*	236.940 (1)	55.582 (1)	1,729.492 (1)
SS_{group} *(df)*	928.490 (1)**	150.470 (1)*	693.174 (1)
SS_{block} *(df)*	564.676 (3)*	581.946 (3)**	1,865.060 (3)*
$SS_{group \times block}$ *(df)*	1,893.588 (3)**	88.137 (3)	559.706 (3)
SS_{error} *(df)*	3,915.378 (48)	2,214.024 (48)	12,451.055 (48)
SS_{TOTAL} *(df)*	7,370.000 (56)	3,035.000 (56)	16,743.000 (56)
F (p) for group x block effect	7.738 (<.001**)	0.637 (.595)	0.719 (.545)

Notes: * $p<.10$ ** $p<.05$.

Table 21.6 Mean Change in Suspicious Person Calls for Service by Block and Group (JCE)

BLOCK	TREATMENT MEAN (SD)	CONTROL MEAN (SD)
Very high (n=5)	–4.20 (6.979)	18.40 (19.995)
High (n=4)	–12.00 (8.246)	8.25 (5.188)
Moderate (n=13)	3.15 (9.353)	5.69 (5.991)
Low (n=6)	4.17 (6.145)	–5.33 (5.785)
Total (n=28)	–.11 (9.689)	5.96 (11.924)

In Table 21.5 we report the results using this approach. The addition of another term has again reduced the error variance. But more importantly, given our discussion, we now have an answer to the question of whether the treatment differs in effect across the groups. The significance statistic for the interaction between group and block is significant only for suspicious persons. Accordingly, we now have an experimental result that tells us that the effect of treatment on suspicious person calls for service is conditioned by the level of crime in the baseline year (i.e., the blocking factor). Table 21.6 shows these results producing the effect of treatment for each of the four blocks for suspicious persons. The table suggests that the effect is consistent in three of the four blocks representing higher crime rates and that the overall effect of the treatment is much larger in the highest rate blocks. However, in the low rate blocks, which represented much less serious crime areas, the observed treatment effect is in the opposite direction. These results would likely lead the researchers to recommend the hot spot policing strategy tested in the Jersey City study for higher rate drug market hot spots, at least in regard to influencing suspicious persons. It might also raise questions regarding why the treatment did not have the desired effect in the lower crime areas.

Some scholars suggest that interaction terms can be added without a block randomized design.[24] This approach is based on the assumption we noted earlier that randomization is likely to lead to balance between the groups on characteristics that might be measured. The problem of course is, as we noted earlier, that there may be chance relationships between variables in your observed data and this might lead to spurious results. It seems reasonable to us that the inclusion of interaction terms without block randomization requires very large samples. Only in such cases can you rely on randomization providing equivalence across a large number of measures. We recommend samples larger than 300 per group before examining interactions in the data without block randomization. Moreover, the researchers should examine whether the data are balanced (and thus there is no relationship between treatment and covariate) in each specific case. With smaller samples, this approach should be carried out with caution, and the interactions observed should be identified at the outset. In any event, only a discrete number of interactions should be observed.

Chapter Summary

Randomized experiments provide higher levels of internal validity than observational studies in terms of determining the impacts of a treatment or an intervention. Through the process of determining an eligibility pool, randomizing the eligible participants or units, and assigning them to treatment or control groups, researchers can better deal with the problem of confounding in posttest measures relevant to the dependent variable of interest. Randomized experiments have the highest possible internal validity as they allow us to assume that confounding causes of the dependent variable are not a concern—this since treatment has been allocated randomly and we can assume that possible confounding factors are not systematically related to treatment.

For concerns of statistical power related to smaller sample experiments, blocked randomized trials can maximize equivalence of experimental studies. Another way of increasing the statistical power in experimental studies is with the inclusion of covariates; however, covariates should be used cautiously as they can allow the researcher to manipulate the results of the study.

One criticism of experimental research is that in its simplest form it is unable to examine contextual factors that may influence treatments or outcomes. Although many experimental studies in criminology have not looked at interaction effects between treatments and other factors that should not be taken to mean that interactions cannot be observed with experimental designs. The best way to examine such interactions is to define them at the outset and use block randomization techniques.

[24] See Barak Ariel and David Farrington, "Block Randomized Trials," In Alex R. Piquero and David Weisburd (eds.), *Handbook of Quantitative Criminology*, (New York: Springer, 2010: 437–454).

Key Terms

alternation A type of quasi-random assignment in which researchers assign every other case to one particular group.

block randomization A type of randomization whereby cases are first sorted into like groups and then afterwards randomly allocated into treatment and control conditions.

confounding factors Variables associated with treatments and/or outcomes that can bias overall results if not controlled for statistically.

control group The group that eligible cases are randomly assigned to which does not receive the treatment or the intervention being evaluated. In many criminological experiments the control group may receive existing interventions in contrast to the innovative treatment.

eligibility pool Participants or units that are eligible for an experiment.

group allocation In criminological experiments, eligible cases are randomly assigned to two or more groups—typically treatment or control.

internal validity Whether the research design has allowed for the impact of the intervention or the treatment to be clearly distinguished from other factors.

posttest measure Analyses conducted by the researcher to determine if the intervention had any impact on the outcome measures of interest.

randomization The process of randomly assigning members from the pool of eligible participants or units to the study conditions—often a treatment group and a control or a comparison group.

treatment group One group that eligible cases are randomly assigned to which receives the treatment or the intervention being evaluated.

Symbols and Formulas

To compute the regression coefficient b for a treatment variable (Tr) controlling for a confounding factor (CC):

$$b_{Tr} = \left(\frac{r_{Y_i Tr} - \left(r_{Y_i CC} r_{TrCC} \right)}{1 - r^2_{TrCC}} \right) \left(\frac{S_Y}{S_{Tr}} \right)$$

With the basic assumption that we have enough knowledge to create equivalence of units in the treatment and control conditions:

$$b_{Tr} = \left(r_{Y_i Tr} \right) \left(\frac{S_Y}{S_{Tr}} \right)$$

T-value for dependent samples:

$$t = \frac{\bar{X}_d - \mu_d}{\sqrt{\frac{s_d^2}{df}}}$$

T-value for independent samples:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{df}} \sqrt{\frac{N_1 + N_2}{N_1 N_2}}}$$

Sum of squares for a simple or a naïve randomized experiment:

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{error}$$

With the introduction of a blocking factor:

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{block} + SS_{error}$$

With the introduction of covariates:

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{covariates} + SS_{error}$$

Accounting for interaction between the treatment and block:

$$SS_{total} = SS_{intercept} + SS_{group} + SS_{block} + SS_{group \times block} + SS_{error}$$

Exercises

21.1 Danny randomly allocates 30 students to either a treatment group that receives new instructional program or a control group that receives standard instruction. After randomization, he compares the characteristics of the treatment and control groups and finds that the treatment group is significantly different from the control group in two characteristics (age, reading level).

a. Do these significant differences between the treatment and control group indicate that randomization failed?

b. What about Danny's sample might explain these significant differences?

21.2 Mark finds in a bivariate regression analysis that a drug treatment program has a significant impact on reducing the likelihood that patients will relapse. The standardized beta coefficient is −0.50. Mark concludes that

the drug treatment program is effective. Brent, however, argues that Mark's results are confounded because he did not account for patients' level of motivation. Brent notes that motivation and likelihood of relapse are highly related ($r = 0.50$). He reruns the regression results controlling for the level of motivation and finds that the impact of treatment has declined. The new standardized beta coefficient is −0.125.

a. Diagram the impact of treatment on likelihood of relapse based on Mark's initial result.

b. Diagram the impact of treatment on likelihood of relapse using Brent's analyses.

c. What is the level of bias Mark has introduced by not including this confounder? What is the estimated r between the level of motivation and treatment?

21.3 Darcy wants to test the effectiveness of a new police training program on domestic violence. She identifies the officers with the least knowledge of domestic violence and administers the training to this group because she believes that it will be most worthwhile since they have the most to learn. She tests this group on domestic violence knowledge before and after the training. She also tests a comparison group of officers who did not receive the training. She finds a major jump in knowledge in the trained officers compared to the non-trained officers and concludes that her training program was effective.

a. Are Darcy's conclusions warranted? Are there any threats to internal validity in her research design?

b. Design an alternative study to test the effectiveness of the training program that has a higher level of internal validity than Darcy's study.

21.4 Adrian is designing a randomized trial to examine the effectiveness of a program designed to reduce recidivism in offenders. He has a sample of 200 prisoners that will all be released from prison on the same day and can be randomly allocated to a treatment group receiving the program or a control group that does not receive the program.

a. If Adrian uses a naïve randomization procedure, how many prisoners will be in each group? What will be the total degrees of freedom for the research design?

b. If Adrian uses a fully blocked randomization procedure, how many pairs of prisoners will be randomized? What will be the total degrees of freedom for the research design?

c. If Adrian wants to use a partially blocked randomization procedure, what might be one prisoner characteristic he uses to create statistical blocks? What are the statistical consequences if this prisoner characteristic does not end up being related to the effectiveness of the program?

21.5 Logan is reexamining data from a policing experiment to assess whether using blocking provided a statistical benefit. Results from a "treatment-only" model that did not include the blocking factor and a "treatment and block model" that did include the blocking factor are provided below.

UNIVARIATE ANALYSIS OF VARIANCE MODELS	CALLS FOR SERVICE
TREATMENT-ONLY MODEL	
$SS_{intercept}$ *(df)*	260.50 (1)
SS_{group} *(df)*	300.65 (1) *
SS_{error} *(df)*	10,875.75 (80)
p-value for group effect	.038*
Treatment and block model	
$SS_{intercept}$*(df)*	500.25 (1)
SS_{group} *(df)*	300.65 (1) *
SS_{block} *(df)*	3,614.44 (5)*
SS_{error} *(df)*	7,021.56 (75)
p-value for group effect	.022*

*$p < .05$

a. What is the total sum of squares in each model? Are these values the same? Why or why not?

b. Why is the SS_{group} the same in both models?

c. How many total blocks were used in the treatment and block model?

d. Did blocking provide a statistical benefit?

21.6 Sharon is analyzing data from a large randomized trial of the impact of after-school programs on juvenile delinquency. After completing the experiment she has been considering adding a number of different covariates to her overall analysis to minimize the error and improve her ability to identify a treatment effect.

a. Do you have any concerns about the approach Sharon is taking to analyzing the experimental data? If so, what would be a better approach?

b. If Sharon has chosen good covariates, what should happen to the SS_{total} in the model? What should happen to the SS_{error}? What should happen to the SS_{group}?

c. With the large sample size, the statistical power in Sharon's experiment is estimated to be about 0.9. Does this affect whether she should consider using covariates?

21.7 Refer to Table 21.6 that provides the mean change in suspicious person calls for service by block and group.

a. This chapter described what the differences by block suggested for the effectiveness of the treatment across the statistical blocks. What do the results for the control group show?

b. What does the total change versus the change in each block suggest about the importance of block by group interaction effects?