CHAPTER 20

An Introduction to Experimental Criminology

LAWRENCE W. SHERMAN

Experimental criminology is scientific knowledge about crime and justice discovered from random assignment of different conditions in large field tests. This method is the preferred way to estimate the average effects of one variable on another, holding all other variables constant (Campbell and Stanley 1963; Cook and Campbell 1979). While the experimental method is not intended to answer all the research questions in criminology, it can be used far more often than most criminologists assume (Federal Judicial Center, 1981). Opportunities are particularly promising in partnership with criminal justice agencies.

The highest and best use of experimental criminology is to develop and test theoretically coherent ideas about reducing harm (Sherman 2006, 2007), rather than just "evaluating" government programs. Those tests, in turn, can help to accumulate an integrated body of grounded theory (Glaser and Strauss 1967) in which experimental evidence plays a crucial role. When properly executed, randomized field experiments provide the ideal tests of theories about both the prevention and causation of crime.

The advantages of experimental methods help explain why this branch of criminology is growing rapidly (Farrington 1983, Farrington and Welsh 2005), with its first journal (*Journal of Experimental Criminology)* and a separate Division of Experimental Criminology of the American Society of Criminology established just since 2005. Yet these advantages depend entirely on the capability of the experimenters to insure success in achieving the many necessary elements of an unbiased comparison. Many, if not most, randomized field experiments in criminology suffer flaws that could have been avoided with better planning. The lack of such planning, in turn, may be due to the scant attention paid to field experiments in research methods' texts and courses. Even skilled, senior researchers can make basic mistakes when conducting field experiments, since experiments require a very different set of skills and methods than the "normal science" of observational criminology. As in any complex work, the value of 10,000 hours of practice can make an enormous difference in its success (Gladwell 2008).

The goal of this chapter is to help its readers improve the design and conduct of criminological experiments. The chapter's method is to describe the necessary steps and preferred decisions in planning, conducting, completing, analyzing, reporting, and synthesizing high-quality randomized controlled trials (RCTs) in criminology. The evidence for the

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_20, © Springer Science + Business Media, LLC 2010

chapter comes from both the authors' experience in conducting such experiments, as well as from published literature: on statistics (especially biostatistics), systematic and nonsystematic reviews of research in criminology, and reports on experiments in criminology. The chapter defines the concept of "experimental method" broadly, so as to embrace the social processes, infrastructure, and personal skills needed to conduct field experiments successfully.

Metaphors for experiments. The success of experimental criminology may depend on choosing the right metaphor. For researchers who are used to finding and analyzing an existing data set, the most tempting metaphor may be a recipe for "baking": mix ingredients, put in the oven, remove data when ready, and then analyze. No metaphor could be further from the mark. More appropriate metaphors for experiments might be found in raising children, governing a small village, or chairing a university department. The most useful metaphor, however, is constructing a building. Construction is a job, like all group life, that requires careful attention every day. But it also has a set of finite features that are much like a randomized experiment: finding money, drawing up and negotiating blueprints, buying land, hiring contractors, and then putting the building up in a controlled sequence of steps.

This chapter is organized around the sequence of steps required to complete a successful field experiment in crime and justice. These steps are conceptual, social, and methodological. The recurrent metaphor of constructing a building helps to illustrate the order of steps to take for best results. The property must be chosen before design can even begin. A building's foundation must be laid before the frame goes up, and the frame must be finished before the roof is installed. Some steps may be arranged in less rigid order, such as whether to install a staircase banister before or after installing the roof. Yet even in that case, evidence can indicate a preferred sequence: the lack of banisters may increase the risk of injury to workers who fall off stairs, for example. What is a "successful" building may thus be seen to have more dimensions than whether the building remains standing (or falls down – as parts of many cathedrals have done).

The steps presented in this introductory chapter begin with the intellectual property of every experiment: formulating the research question. We also consider the social foundation of a randomized experiment: choosing and developing the field research "station," as the science of agriculture describes it. In criminology, a field station can comprise many different kinds of settings, but all of them must (by definition) have some agents actively delivering some treatment(s) that will have hypothesized effects on crime or justice outcomes. Once a partnership is established between these agents and the experimenters (who could even be the same people), the next step is developing a research "blueprint" for building the experiment, known as the research protocol. This step may include a "dry run" of treating cases that will not be included in the final experiment, because the evidence from such pretest cases may prompt a change in experimental design. Other decisions must be made at that stage to address the issues described in the following sections, including the requirements of the CONSORT statement about reporting randomized trials in all fields of study.

Once a protocol is agreed and approved, the experimenters (like builders) must find and "contract" with a wide range of agents and others to construct and sustain the experiment in the most favorable way possible. Sustainability depends on well-planned and responsive management of the experiment, covering each of the following steps: supplying cases, screening for eligibility, randomly assigning treatments, delivering treatments consistently, measuring treatments delivered, and measuring outcomes.

When and if all these steps are completed, the experiment will be ready for analysis. The basic principles of experimental analysis are partly addressed in the general literature on data analysis, but several principles unique to experiments are widely violated. The chapter briefly maps out those principles and the arguments for and against fundamentally different analytic approaches in experimental criminology (EC). A separate section addresses the communication of the results of experiments to both professional and lay audiences. The penultimate section addresses the synthesis of experiments into more general knowledge, including the benefits of designing experiments as prospective meta-analyses, or "REX-Nets," and some examples of them in EC. The chapter concludes with reflections on the personal skills, training, and apprenticeships needed to practice EC, in light of the many tasks required to complete experiments successfully.

The chapter is intentionally "front-loaded," concentrating far more on the early stages of an experiment than on its final phases. The latter are far more extensively covered in standard research methods texts. Sadly, even the best research methods cannot make up for failings in research design. Because those failings are best avoided by better strategic planning at the beginning of an experiment, this chapter invests most where it may do the most good. Appendix 1 distills much of the planning needed into a protocol, which we have named the Crim-PORT or Criminological Protocol for Operating Randomized Trials. We invite anyone who is launching a randomized trial in experimental criminology to use the Crim-PORT in designing it. Even better, we invite them to register their trial protocols at the Cambridge Criminology Register of Randomized Controlled Trials (http://www.crim.cam.ac.uk/experiments).

INTELLECTUAL PROPERTY: FORMULATING THE RESEARCH QUESTION

> The mere formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill.
>
> —Albert Einstein and Leopold Infeld
> ([1938] 1971: 92)

What is "merely" a matter of experimental skill is the subject of this entire chapter and of thousands of books. The subject of the present section of this mere chapter is the skill required in formulating testable questions. That skill begins with a conceptual appreciation of what questions are theoretically important. But it also includes a technical appreciation of what hypotheses are testable. Finally, it includes a utilitarian conception of how important a question is from a cost–benefit perspective. These considerations provide the basis for laying the intellectual foundations of a great experiment.

Great experiments in criminology are arguably based on these three criteria:

1. They test theoretically central hypotheses.
2. They eliminate as many competing explanations as possible.
3. They show one intervention to be far more cost effective than others.

Putting these three criteria together in the formulation of an experimental research question may seem to be more a matter of "art" than of science. But such a judgment would demean the importance of intuition, inspiration, and insight in science, as in many fields involving complex decisions (Gladwell 2005). What will always distinguish great from routine science is the capacity of a given contribution to make a major leap forward in understanding (with theory), intervention (with public benefits), or both.

Testing Theoretically Central Hypotheses

Better theory is the initial goal of all science; public benefit is arguably the ultimate value of any theory. Public benefits may (or may not) follow from better theory, sooner or later – and sometimes much, much later. James Lind's experiments in the prevention and treatment of scurvy on long sea voyages led to the British Navy issuing citrus fruit to all their sailors – hence the term "limies" – which saved thousands of lives. Yet the time between research results and policy change was substantial: it took over four decades for the Navy to act on the experimental results (Tröhler 2003). Ignaz Semmelweis showed in a nearly randomized experiment that gynecologists could reduce death in childbirth by washing their hands before examining new mothers, but half a century elapsed before the findings were widely accepted in the medical profession (Loudon 2002).

Neither Lind nor Semmelweis conducted experiments that were seen as critically important to a current theoretical debate, which may enhance the speed with which the findings are accepted (Tilley 2009). But that does not mean that the experiments failed to make major contributions to theory. To the contrary, the fact that Lind's and Semmelweis's experiments are still celebrated today shows just how important they were in the long run, theoretically as well as in terms of public benefit. Lind laid the foundation for theories of the immune system, while Semmelweis offered crucial evidence for the germ theory of disease.

In criminology, the deterrence doctrine has been central to both theory and policy for over two centuries. Yet few unbiased tests of that doctrine have been conducted. The lack of definitive experiments has limited the development of deterrence as a formal theory that specifies the conditions under which punishment or its threat prevents crime (Gibbs 1975). Great experiments in criminology have begun to advance that theory in predicting effects of legal sanctions on the criminal behavior of punished individuals, their communities, and their nations (Sherman 1993). They could also create great public benefit, limiting the harm of punishment under guidance of experimentally supported theory.

The promise of experimental criminology (EC) is equally great for understanding other social responses to crime, such as the engagement of schools and families in the prevention of serious delinquency (Multisite Violence Prevention Project 2008). Large-scale, multidisciplinary research on such questions can identify differential effects of similar responses on different kinds of people or communities. Of special theoretical importance is any finding of opposite effects of identical programs on different kinds of people, especially low-risk vs. high-risk persons (Erwin 1986; Sherman and Smith 1992; Hanley 2006; Multisite Prevention Project 2008). Experiments designed to test for such differential effects help advance theory by showing the interactions among the multiple causes of crime.

When experimental criminologists are asked to help evaluate innovations or programs intended to reduce crime or injustice, they are rarely asked if the innovations make any sense in terms of theory. Rather, the question is baldly put to them: does this program work? Often the best response is not empirical, but theoretical: why should it work? What is the theory of cause and effect implicit in the design of the program, and what prior evidence (if any) is consistent with that theory?

Experimentalists can do the most good for science when they are the most focused on the theoretical implications of their experiments. Evaluation research methods often assume – or require – a passive role for an "evaluator" in the design of a program, in order to keep the evaluation "independent" and free from conflict of interest (Eisner 2009). Yet society's limited resources for research are better served when experimental criminologists actively help to reshape a program before it is tested, in order to make a better contribution to theory. That is a role many experimenters have played without deriving any financial benefit from the design or success of the program (Sherman and Strang 2010).

The author was fortunate to have been asked to design repeated RCTs (14 completed) of the globally influential theory of reintegrative shaming (Braithwaite 1989). The first four of these experiments (Sherman et al. 2000) allowed more detailed measures of the constituent elements of the theory, which led to the theory's revision (Braithwaite 2002). More important for experimental criminology, the theory itself was adapted into a program (restorative justice) supported by a social movement that has welcomed the RCT evidence as a tool for helping to transform contemporary justice. It has also become a prime example of how RCTs can generate evidence on cost effectiveness (Shapland et al. 2008).

In their most theoretical posture, experimental criminologists have designed field experiments for the sole purpose of testing theories of crime *causation*. Farrington and Knight (1980), for example, randomly assigned the characteristics of potential crime victims to determine the effects of those characteristics on decisions of potential offenders to commit crime in an anonymous urban setting. Their findings put substantial empirical flesh on the bones of routine activities theory, which hypothesizes that a potential victim's "suitability" is a central condition for crime to occur.

There are always limits, of course, to how well theory can be tested in collaboration with governmental agencies wielding the power to manipulate variables of theoretical interest (Farrington 2003). Ariel (2009), for example, proposed theoretically coherent versions of letters that a tax agency could send to taxpayers. After negotiations for a large-sample experiment were concluded, the experiment had a sample of over 16,000 cases randomly assigned to receive different letters, with direct measurement of taxes subsequently paid. The content of the letters, however, was altered substantially by the agency for administrative reasons, thus limiting somewhat the theoretical payoff from an experiment of great policy significance. The experiments were an excellent test of the specific text of the letters. But because the government did not allow the text to be stated in a way that would tap the key theoretical questions about compliance with law, the results could not be directly incorporated into a grounded theory of tax compliance or broader theory of social control.

Whatever the practical limits to theory testing may be in an experimental design, the only way to find them is to push for as much theoretical benefit as possible. As Ariel (2008) observed, his experiment consisted of "three years of negotiation and one day of random assignment." The value of clear testing of a theory justified every day of those 3 years as a fight for strong theoretical implications of the results. For in theory as in policy, unambiguous results have the capacity to move discussion forward. It is the capacity to limit ambiguity by eliminating competing explanations that makes EC so important to criminological theory.

Eliminating Competing Explanations

> "...when you have excluded the impossible, whatever remains, however improbable, must be the truth."
>
> —Sherlock Holmes, *The Adventure of the Beryl Coronet*, by Arthur Conan Doyle[1]

[1] See http://www.online-literature.com/doyle/adventures_sherlock/11/, downloaded on July 24, 2009.

The reason that experimental criminology can make major advances in theory is the strong internal validity of the RCT design (Campbell and Stanley (1963). Internal validity is the extent to which a research design can eliminate competing explanations of a correlation. The more "plausible rival hypotheses" about a correlation that a study can eliminate, the more likely it becomes that the surviving explanation is the "true" cause of the observed correlation. Assume, for example, that you are taller than your father and that you drank more orange juice while growing up than he did. A theory to explain your height difference might then be that orange juice caused it – or at least some of it. But many other factors could also have caused your difference in height. The only way we can be confident about a causal impact of orange juice intake is to eliminate all (or most) other possible explanations for a height difference. While it is arguably impossible to assess causation just for you and your father – i.e., in a single, anecdotal comparison – we can learn, on average, what difference orange juice intake may cause across a large group of people.

One way to isolate the independent effects of orange juice – perhaps the best way – would be a prospective, longitudinal experiment that manipulated the amount of orange juice a large sample of people drank while they were growing up. To use this ethically unimaginable example as a hypothetical, we could randomly assign 2,000 children into two groups. We could pay 1,000 of them to drink 12 ounces of free orange juice daily and pay the other 1,000 *not* to drink more than 3 ounces of orange juice a day. Everything else about growing up would remain the same – varying from one child to the next, but with the averages and percentages of almost everything from their television watching to their cigarette smoking being almost identical in each group. Why? *Because random assignment makes it so.*

The "magic" of random assignment is that, with large enough samples, it almost always yields similar distributions in two different groups of the potential causes of any future behavior by the members of those groups. This should mean that the only average difference between the groups is the one that the experimenter has *independently* manipulated: independent, that is, of the normal preferences and habits in drinking orange juice across 2,000 children. In our orange juice example, any difference in average height between the two groups (the *dependent variable)* could then be attributed to the difference in (measured) intake of orange juice, on average, between the two groups (the *independent* variable).

EC thus shares the common distinction between experimental and nonexperimental branches of all fields of science: it tests causation by *systematically altering* variables in the units of study, as well as by *observing* those units over time. All empirical science systematically observes phenomena, but only experiments intentionally (or "independently") manipulate one or more of those variables. The power to independently modify variables under observation is so important that some statisticians have suggested that it is essential to understanding cause and effect (Salsburg 2001: 181–194), using the adage:

"No causation without manipulation."

Criminology graduate students are usually trained, improperly, to "control" for mere correlations using statistical equations in multivariate analysis in the hope that observational controls will identify causal pathways. This may imply that they are "manipulating" the variables with statistics.

They are not.

They cannot.

Manipulation means actually *changing* a factor in real life, *doing* something to someone or something. Merely selecting cases or writing equations is not manipulation. Nor does statistical "control" demonstrate causation with the same strong internal validity as random assignment to experimental and control groups.

The reason that observations alone fail to eliminate competing explanations is that they require data analysts to be too smart and too lucky. Unless analysts are both *smart* enough to think up (or "specify") every variable that needs to be held constant and *lucky* enough to have a data set in which all possible conditions of all relevant variables have enough cases for analysis of the primary causal hypothesis, statistical controls are not enough. Causation cannot be *strongly* inferred without being sure that you were both smart and lucky. Sadly, there is little way to tell without stronger research designs. Not all of these designs entail randomized experiments (see other chapters in this volume), but they all entail more than multiple regression.

Natural and Randomized Experiments. Sometimes this problem can be largely solved by an experiment of nature, in which the observer must merely document a process that was independently manipulated by factors outside the phenomenon under study (or an "instrumental variable"; see Angrist et al. 1996; Angrist 2006). That is exactly what David Kirk (2009) did with Hurricane Katrina, as a test of the well-developed and data-grounded theory that criminals are more likely to desist from crime if they succeed in "knifing off" their contact with the social networks that have supported and encouraged them in criminal activity (Laub and Sampson 2003). This theory had been tested by prospective, life-course measurement in a number of settings prior to Hurricane Katrina. But none of those tests could rule out competing explanations that were also correlated with the result, such as making more use of the educational benefits made available to World War II veterans.

It was only the hurricane that could provide an *independent* manipulation that "knifed off" prior friendships. The manipulation was independent – or random, in the statistical sense – because the hurricane was not influenced by social factors related to crime. What Kirk (2009) knew was that some New Orleans residents who came out of prison could not return home after Hurricane Katrina wiped out some, but not all, high crime neighborhoods. When he compared those "homeless" offenders to those who *could* return home, he found the now-homeless offenders were much less likely to be sent back to prison. In general, those now-homeless who resettled farthest from their last address before entering prison were the least likely to be reincarcerated for a new crime.

This evidence is more convincing than previous tests because the hurricane was disconnected from other factors embedded in the life course of the offenders. But did it eliminate all competing explanations? Arguably not, as Kirk (2009) carefully observes. Nor would anyone build a crime prevention policy that relied upon hurricanes.

More important is the theoretical nature of the intervention. Moving people from one community to another is itself theoretically ambiguous. Even if the intervention were designed as an RCT, it would not be easy to eliminate all competing explanations. One competitor, for example, is that when offenders move to different cities (or even neighborhoods) they are not as well known to police. Nor do the usual police informants know them or hear rumors about their criminal activities. Thus rather than committing fewer crimes (desisting) because they have "knifed off" their prior relationships, a competing explanation for lower reincarceration rates is a detection hypothesis. Offenders who change locations, it could be argued, commit just as many or more crimes than they did in their old neighborhood. They are simply less likely to be detected, arrested, convicted, and reincarcerated where they are unknown than where they grew up.

The point of considering the Kirk (2009) Katrina study, in this context, is to illustrate the process of designing a great experiment in criminology. Kirk's study was a brilliant formulation of a research question about a natural experiment. What it implies for EC is the development and testing of a *program* designed on the *theory* associated with "knifing off" in the life course (Laub and Sampson 2003). This could involve subsidizing offenders' moves to new (and distant) communities, with more support and integration into a community. Unlike a natural experiment that destroyed old neighborhoods, it would require measurement of the rate at which offenders give up on the new community and move back to their old one (measurement of treatment condition). It could also require measurement of the extent to which offenders stay in touch with old contacts.

Whatever an RCT design does to build on Kirk's natural experiment, its success will depend on its ability to rule out, or at least test, competing explanations. Thus if reentering offenders have less reincarceration when they are randomly assigned to move far away, the key question will be whether they are really desisting from crime. The experiment must therefore build in one or several strong tests of the detection hypothesis. Self-reported offending, DNA checks of crime scenes, or even efforts to make local police aware that the offenders have moved in – these and other strategies could help address a competing explanation.

Competing Explanations of a Good Result. In the absence of such tests of competing explanations, even positive results from one or more RCTs may fail to convince the prime audience for the experiments: theorists and policymakers. A prime example of this is the hot spots policing experiments. Ever since Sherman and Weisburd (1995) demonstrated that extra patrols suppressed crime and disorder in 55 Minneapolis hot spots (compared to 55 controls), some police and crime theorists have remained skeptical of the result. They point to the competing explanation that hot spot policing simply "pushes crime around the corner."

In a series of subsequent studies, Weisburd refuted the displacement hypothesis by demonstrating that crime actually declines in the immediate vicinity of areas targeted with extra patrol – a finding for which he won the 2010 Stockholm Prize in Criminology. Yet many police and criminologists remain unconvinced. Weisburd has even demonstrated that once drug markets were shut down in Jersey City, they did not pop up in other parts of the city. But skeptics suggested that the markets could have moved to another nearby jurisdiction.

One reason for continued skepticism is the impact of extra policing on the individual offenders who were committing crime in the hot spots. No published evidence to date has tested the hypothesis that when crime goes down in one location the offenders simply move to other locations, perhaps at some distance, within the same metropolitan area. That is one reason why the Greater Manchester Police are now working with the present authors to conduct a new hot spots patrol experiment that would track offenders' arrests across the 500 square-mile area of their jurisdiction, in contrast to the 15 square miles of land in Jersey City (or even the 300 square miles of land in New York City). By testing for displacement across such a broad catchment area, a Greater Manchester experiment could falsify the individual offender displacement hypothesis more convincingly than previous tests – assuming it finds a reduction in offender frequencies of arrest or convictions. Such a finding would be predicted by routine activities theory; its converse would falsify the theory and support the "hydraulic pressure" displacement espoused by so many police and theorists.

No matter what the results are, the point of formulating the right research questions is to *anticipate* competing explanations for any test of a theory. As the Katrina and Manchester examples suggest, anticipating rival explanations may require enormous work in finding the right places to do the experiment. Yet that work is well worth the investment. The more competing explanations an RCT design can rule out, the more convincing the results will be to Sherlock Holmes and his numerous followers.

Demonstrating Cost Effectiveness

The greatest advertisement for EC is research demonstrating the cost effectiveness of one strategy over another. In recent years, rising interest in this principle alone has done more to encourage evidence-based government than any other. From the creation of the Washington State Public Policy Institute (http://www.wsipp.wa.gov/) to President Obama's 2009 Inaugural address and from the UK's National Institute for Health and Clinical Excellence (http://www.nice.org.uk/) cost-effectiveness standards for medical treatments to the "evidence help desk" established by the Bush administration in the US Office of Management and Budget with support from the Jerry Lee Foundation (http://evidencebasedpolicy.org/wordpress/), the cost effectiveness of government programs is attracting more analysis than ever. At the same time, these developments favor the use of experimental evidence over research designs with less internal validity.

These developments notwithstanding, EC is still more focused on theory and explanation than it is on cost effectiveness. The plea by Welsh et al. (2001) that experimentalists should monetize the benefits found in crime prevention experiments has barely been heeded. A review of articles in the *Journal of Experimental Criminology* since its founding in 2005 shows that very few authors report their effects in financial terms – including the present author! Instead, there is a clear emphasis on reporting effect sizes as standardized mean differences or Cohen's D. While that statistic offers the mathematical appeal of making effect sizes more comparable across experiments, it is impossible to derive much policy or political appeal from purely mathematical formulae. This, however, is *mostly* matter of presentation and not a fundamental problem of shaping a research question.

The framing of research questions for great experiments in criminology must do two things to capture cost effectiveness. One is that experiments must be planned to measure the costs of delivering programs, both in a start-up (developmental) phase and in a "rollout" model with perhaps more efficiencies from mass production. Without gathering measures of personnel time, travel costs, or equipment in delivering a treatment during the course of the experiment, it is often impossible to compare costs of treatment and control conditions. The same is true for the dependent variables. Offense-specific measures of crime, as distinct from counts of crimes or even mere prevalence of repeat offending, can say much more about the benefits of crime reduction. Yet it is common for many experiments to limit their measurement to counting any crimes rather than recording the *types* of crime committed. Once those types are known, a far more precise – if not perfect – assessment of the costs of crimes prevented can be calculated. These calculations can use recent average costs tracked in national samples. The costs of crime are increasingly available in standardized estimates, such as the UK's Home Office (2005) calculations for England and Wales.

The second requirement is far more important and less technical than the first. The requirement is simple: *test the most expensive strategies*. Frame experiments in ways that address major questions of cost and harm, where the payoff of the research can become a big "win" for experimental criminology. Showing how prison populations can be reduced or prisons actually closed (at US$40,000 or more per year per inmate), for example, without increasing crime, could appeal to all shades of politics. Showing how murders can be prevented, at US $2 million per murder (Home Office 2005), might have far greater cost-effectiveness value than showing how auto theft can be reduced. In short, when framing the research question for an experiment, look for ways to test not just important theories, but ways to reduce major economic and social costs allocated to crime responses.

One way to do this is for experimental populations to be selected on the basis of Pareto's principle of the "power few" (Sherman 2007). By focusing on the small proportion of offenders (or of victims or street corners) that produce the most crime, experiments have the greatest potential to demonstrate large differences in cost effectiveness of two policy choices. The "million dollar blocks" of big cities generate enough prison inmates to cost US $1 million per year. The high cost of these blocks has become the basis for an entire – but as yet unevaluated – strategy called "justice reinvestment." The strategy illustrates a "power few" approach that could readily become the basis for random assignment of a standard intervention across high-inmate-cost residential street blocks.

SOCIAL FOUNDATION: DEVELOPING A "FIELD STATION"

The least appreciated requirement for doing EC is its social foundation. You cannot (usually) just walk into an agency and propose an experiment. Even if the agency agrees, the experiment is likely to fail without constructing firm social capital at the outset. The foundation of social capital is the set of human relationships and social networks linking what we call the "experimenters" and those who control the resources to be allocated by random assignment. The experimenters are criminologists trained in experimental design. The people who control the resources can be called research partners, such as leaders and middle managers of police agencies, schools, courts, probation agencies, prisons, social service agencies, and parole boards.

This task is getting more complex as "evidence-based" or "data-driven" practices become more fashionable. That fashion has led to in-house hiring of research staff who may actively resist external research collaborations. What was once a bilateral relationship between external experimenters and internal operating personnel is more often a trilateral relationship between external and internal researchers and the operating personnel. Sometimes this may simply block experiments from ever getting started. In other cases, the experiments proceed, but with structural tensions involving the research roles more than the operating roles.

Even the bilateral relationships of operators and experimenters can deteriorate into a mutually suspicious, us-vs.-them antagonism. This has been especially likely when experimenters were brought into a relationship with partners as "evaluators," when an evaluation was required by a third party, such as a program funder. Increasingly, however, EC has solved this problem by abandoning long-term relationships with *funders* in favor of long-term relationships with operating agencies. These agencies, in turn, are increasingly willing to invest their own funds in the conduct of randomized experiments.

A long-term social foundation for EC can support many more than one experiment. That multitest approach to a single site helps to earn back the heavy initial costs for experimenters in learning how the partner agency operates. It also helps the partner agencies to earn back the time they have invested in teaching the experimenters what they need to learn to do just one experiment; for the same "sunk" costs, the partnership can do 2, 3, 5, or 10 experiments – or more, if they take the long view.

Field Stations in Experimental Field Sciences

The history of experimental field science shows many examples of research centered in what looks much like an indoor laboratory – but with a crucial difference. Any "field" science, by definition, studies questions that cannot be answered in a laboratory. How do infections actually spread in populations of humans or of cows? How cost effective is the addition of fertilizer to crops? How many geese will fly south from Canada to the US next year? Not all of these questions require experimental methods (only the fertilizer test does). But none of them can be answered in a laboratory, at least not directly.

Field research stations have collected various kinds of observational data systematically in the same places for at least 300 years, ever since a network of weather stations was set up across Italy (Bradley and Jones 1992: 144). Experimenters have conducted various experiments in field settings for at least two centuries, at least since Benjamin Franklin went to the belfry of Christ Church in Philadelphia to fly a key from a kite in a thunderstorm to see if he could attract lightning to it (he did).

What was perhaps the first *experimental* field site was the Rothamsted Agricultural Experimental Station established in 1843 by Sir John Bennet Lawes, owner of the estate and its fertilizer factory (also the first such factory in the world). In a 57-year partnership with chemist Joseph Henry Gilbert as the "experimenter," Lawes established a program of "classical experiments" in how to increase crop production, largely through fertilizers. Some of these experiments are still being conducted.[2] Even more important was the intellectual product of the Rothamsted Station at the end of its first century: the classic treatise on the design of RCTs by Sir Ronald Aylmer Fisher (1935).

Many other experimental field stations have been established around the world since 1843. Under the 1862 Morrill Act in the US, every state was entitled to establish a land-grant agricultural school, most of which established experimental field stations to test farming practices on large samples of fields or grazing animals. By the 1950s, hospitals associated with medical schools took on the same character as field stations, linking teaching and research with a large number of clinical RCTs. These "power few" hospitals developed the greatest concentrations of research grants and researcher–practitioners. While it took a half-century from the call by the great medical educator, Sir William Osler, for research universities to "invade the hospitals" (Bliss 1999), the growth of medical experimentation was clearly advanced by its concentration in teaching hospitals with a strong empirical ethos of testing every practice possible.

Field "Stations" in Experimental Criminology

From at least the 1960s, similar concentrations of field experiments have been found in the criminal justice system. The Vera Institute of Justice in New York appears to have been the first center of repeated experiments in criminology, launching its first RCT in October 1961. Comparing money bail to release (without cash) on recognizance in the 1960s, Vera literally changed the world with its conclusion that showed how to reduce jail populations without

[2] See http://www.rothamsted.bbsrc.ac.uk/corporate/Origins.html, downloaded on July 26, 2009.

increasing crime (Ares et al. 1963). Vera went on to treat all five boroughs in New York as one large field station, conducting repeated RCTs (and observational studies) in New York's prisons, courts, prosecutors' offices, and police agencies.

In the same year that Vera began randomized experiments in New York, the California Youth Authority (CYA) began a long program of RCTs on the other side of the US. This "field station" thrived until it was pushed aside by the "just deserts" model of sentencing, which denied the relevance of "what works" in reducing repeat offending as a morally unacceptable question (Palmer and Petrosino 2003). But just as the CYA's research culture was dying out, the Police Foundation launched a series of RCTs and quasiexperiments with police departments around the country, culminating in the random assignment of arrest in 1981–1982 with the unanimous approval of the Minneapolis City Council (Sherman and Berk 1984). Of all the police agencies with which the Police Foundation collaborated, the Kansas City (Mo) PD became the most like a field station, with repeated research grants and quasiexperiments (Sherman 1979; Sherman and Rogan 1995a, b) and with multiple researchers (Sherman 1979; Sherman and Rogan 1995a, b). Other police agencies have played similar roles, including Jersey City and San Diego.

It is arguable whether such "hot spots" of Experimental Criminology should be called "field stations," using the same terminology as experimental agriculture. Many officials prefer the image of a teaching or research hospital, if only because medicine is generally more prestigious than farming. Both fields are quite scientific, however, and the science of policing is spatial in much the same way as the treatment of fields in agriculture. But marketing brand names aside, the concept of a field station where data are recorded and experiments can last for many decades is an explicit vision for how to conduct experiments in criminology. Any investment in EC is best directed to such locations. Anyone leading an experiment is well advised to study the histories of previous such sites, with as many case studies as possible (Sherman 1992, Appendix 2; Sherman and Strang 2010).

The Social Elements of Experiments

The key to holding an experiment together is understanding a cognitive map of its social elements. These elements include (1) the funders, (2) the executive leadership of an operating agency, (3) the mid-level operating liaison person(s), (4) the agents delivering treatments, and (5) (where necessary) the agents providing cases.

1. The *funders* vary widely in their previous experience with EC or with randomized experiments in any field. In some instances, grant managers may know more about randomized experiments than the principal investigators. In other cases, the funding organization has unrealistic expectations of how quickly or feasibly an experiment can be accomplished. The latter problem is especially likely to arise when the funding staff members attempt to design an experiment in the abstract, without being grounded in the daily operations of a particular agency which can carry out an experiment. In one case, a senior prosecutor designed an experiment with 37 criteria for which cases were ineligible – leaving only 15 cases out of over 5,000 that were eligible when the criteria were finally applied in the field. No one in the funding agency ever thought to check the eligibility criteria against field data before the experiment was funded; they did not hesitate, however, to blame the problem on the research team that was awarded the contract to implement (and "evaluate") the funder's design.

Whatever the funder's background, however, a candid and constructive relationship is an insurance policy against the many predictable problems of constructing experiments. Frequent communication and warnings of delays may help pave the way for funder's support for extensions of time or additional resources. These can often be justified on the basis of achievements in the first phases of the experiments, including substantial numbers of cases randomly assigned with a high level of integrity.

2. The *agency executive* is someone who should strongly support the experiment from the first day of planning and then be left alone unless a critical problem requires intervention that no one else can provide. Ideally, the principal investigator should be able to communicate directly with the executive, providing progress reports in writing or in person once or twice a year. This privilege should be used as little as possible, however, especially in very large organizations. Only with a strong preexisting relationship is frequent contact likely to be welcomed. A large program of experiments at any one time would also create a firmer basis for more frequent contacts. Where the executive is frequently replaced during the course of an experiment – as occurred five times in 5 years with our Canberra experiments – a legally binding contract with the agency signed at the outset can keep the experiment going until completion.
3. The *operating liaison* is the most critical player after the principal investigator. A skilled leader such as Lt. Dean Collins in Milwaukee can make a better case for random assignment than any researcher; Collins presented the design to the City Council, which voted unanimously to approve it (Sherman 1992). Lt. Anthony Bacich, also in Milwaukee, selected and led a team of 36 officers who delivered the tightest implementation of random assignment in the history of police experiments. In contrast, the refusal of the Canberra police to designate a permanent liaison for the life of the experiments led to much less consistency and fidelity in the implementation of the experimental protocol.
4. The *agents delivering treatments* should arguably be as few in number as possible and preferably invited to serve in the experiment on a voluntary basis. A labor-intensive experiment that requires all operating personnel to participate, such as the Minneapolis Hot Spots patrol experiment (Sherman and Weisburd 1995), raises far more issues of compliance with random assignment than develop along with a small voluntary team. A smaller team can also become more committed to the experiment by knowing that their names will be on the final report, by attending monthly briefing sessions, by attending a 1–3 day conference at the outset of the experiment, and other means of fostering team spirit. The more effort experimenters can put into such relationships, the more likely the treatments are to be delivered as the protocol specifies.
5. The *agents providing cases* may be entirely different people from those delivering the treatments. Sometimes they may seem almost invisible, such as the dispatcher referring domestic disturbance calls to a small portion of the patrol force who have been designated to conduct the experiment. Ignore them at your peril, since they can choke off the sample size needed for statistical power.

Some will always be more responsive than others, as persistent attempts to win them over may discover. Dispatchers can be visited, briefed, and cosseted. Magistrates' courts clerks in England, however, may smile, attend meetings and luncheons, promise to help, and then refuse to do what even the head of the nation's judiciary asks them to do in person. In that example, their opposition was more to the substance of the intervention being tested than to the idea of experimentation.

No matter what the reason, a failure to persuade those providing cases is the number one source of experiments failing to reach sample size goals. The wary may choose to refuse an experiment in which the sources of cases may oppose their referrals. The bold may choose to forge ahead, as Sarah Bennett, Nova Inkpen, and Dorothy Newbury-Birch did in England. After repeated trial and error with multiple targets for extracting cases prior to sentencing, they discovered a year later that what the courts denied us (same-day notice of guilty pleas), probation departments were willing to supply.

BLUEPRINTS: DECIDING ON THE EXPERIMENTAL PROTOCOL

In the long history of real-world field experimentation, as in building construction, formal blueprints are a recent arrival. Most cathedrals were built without the modern equivalent of architectural plans, yet most are still standing. Medical research rarely required protocols, even when patients were dying (as in the first patient treated with penicillin). Careful records were often kept, tracking fast-moving innovations in finding a better way to carry out tests. Initial plans were often subject to change.

All this has been changed in the modern US university environment of Institutional Review Boards (IRBs), where a formal research protocol is a legal requirement. Any major change in that protocol needs to be reviewed and approved by IRBs in advance of implementation. The negative effect of this requirement on EC is hard to overstate, as we have learned by operating experiments in England without legal requirements for such protocols. We can say that our eight UK experiments would have been virtually impossible to do with constant delays of 2 months or longer, each time we decided to alter a protocol for case referrals or types of offenses eligible. Yet the future is clear: experimental criminology will need to design blueprints and stick to them, absent approval from oversight bodies.

On balance, it can be argued that EC will be substantially improved by wider use of experimental protocols. One reason is that there have been so many RCTs in criminology that either violated good design standards or failed to report fundamental information (Weisburd 1993). The 1996 publication of the first CONSORT statement – the CONsolidated Standards On Reporting of Trials – was prompted by a similar concern in medicine, after surveys documented the common gaps in reporting on medical trials (e.g., Pocock et al. 1987). The elements in the CONSORT statement are an extremely valuable guide to reporting RCTs in any field, not just medicine. But as late as 2009, many leading criminologists and experimenters had never heard of the CONSORT statement (http://www.CONSORT-Statemen.org). Wider knowledge of the statement may lead to more complete reporting of EC. But more important is that it can lead to better planning of experiments with protocols that anticipate these reporting requirements.

The checklist of CONSORT's 22 reporting elements includes the following outline of a final report:

1. Title & Abstract
2. Background
3. Participants
4. Interventions
5. Objectives
6. Outcomes
7. Sample size

8. Randomization – Sequence Generation
9. Randomization – Allocation Concealment
10. Randomization – Implementation
11. Blinding (masking)
12. Statistical Methods
13. Participant Flow
14. Recruitment
15. Baseline Data
16. Numbers Analyzed
17. Outcomes and Estimation
18. Ancillary Analyses
19. Adverse Events
20. Interpretation
21. Generalizability
22. Overall Evidence

Any experimental criminologist reading this list will question the sequence and logic of the items. They may also question the relevance if items such as "blinding" of participants to the treatment they are receiving – a difficult task in criminal justice research when, for example, people are being arrested (or not). But the adoption of at least the CONSORT caseflow diagram could lead to far greater transparency in EC, if not all of its elements (Fig. 20.1).

A flow diagram is an excellent tool for depicting what happened across all randomly assigned cases in an experiment. An illustration of such a diagram above comes from an RCT in Philadelphia comparing low intensity probation with standard intensity (Barnes et al. 2010). It shows the pipeline of cases flowing into the experiment, which was a repeated batch design. It also illustrates what happened to cases after random assignment – showing that all cases were analyzed as if they had been treated as they were randomized, regardless of what actually happened (often due to offender behavior which led to their attrition from probation supervision). Knowing that such a chart is required may insure that the data are gathered as the RCT progresses.

Yet CONSORT alone is not enough. A reporting system does not tell you how to design a protocol for an experiment before it starts. It only tells you what readers need to know after it is finished. There are many essential things you must do to make experiments succeed that the reader has little need to know. These are not only the *formal* elements of a protocol, such as informed consent and statistical power. They are, more importantly, the managerial elements of delivering the experiment as desired. And perhaps because there are so many differences across research fields, there is no consolidated statement for the development of RCT protocols. It is therefore incumbent on experimental criminology to develop its own standard.

For this and other reasons, we include in this chapter the first version of a standard protocol format for experiments in criminology. Appendix 1 lays out the elements of the protocol, with more detailed instructions to be posted on the Web site. The Web site is provided for the dual purpose of (1) helping experimenters to design better trials and (2) providing a public registry where protocols can be posted in advance of a trial starting. Such registries have become essential in medicine to combat the "file-drawer problem": the systematic bias that comes from not reporting tests in which no statistically significant differences were found or differences were found in the opposite direction from where it was anticipated (or hoped).

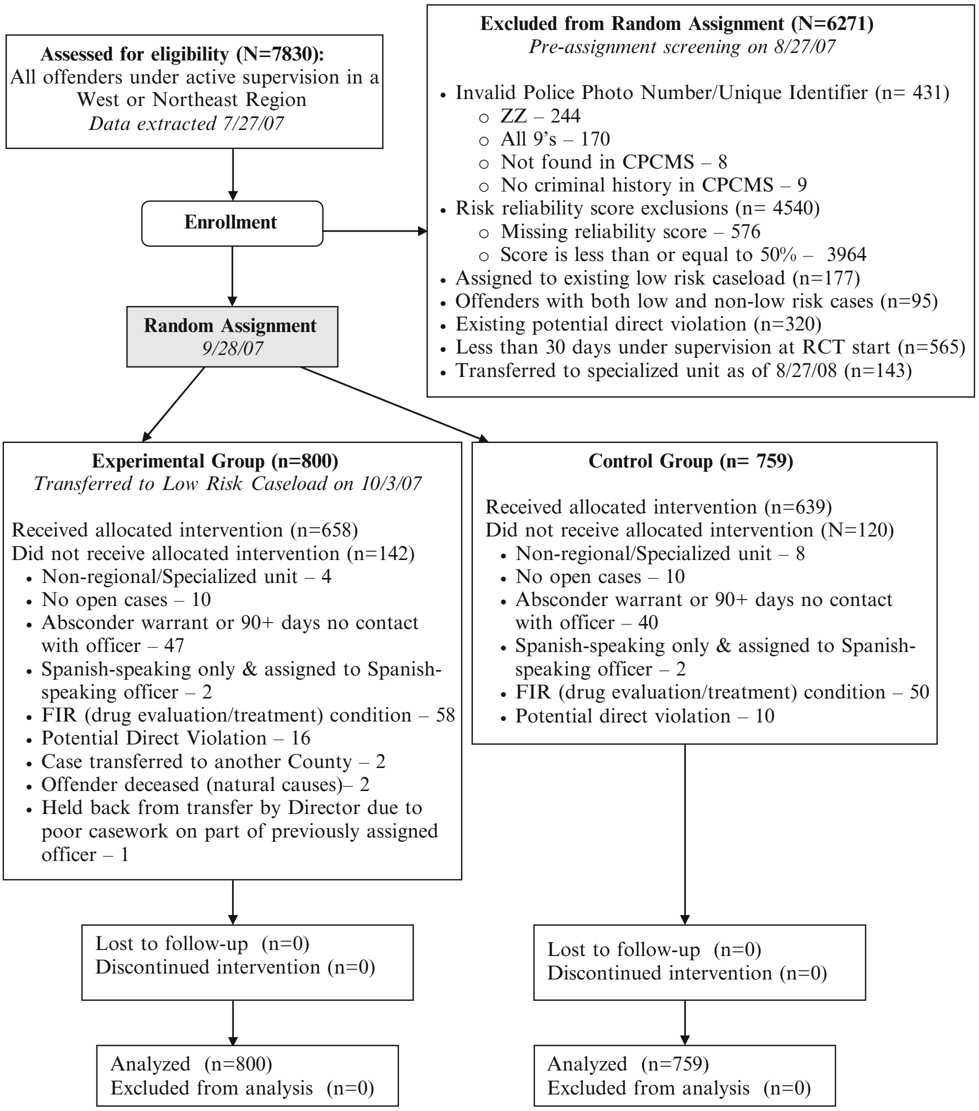
Assessed for eligibility (N=7830):
All offenders under active supervision in a West or Northeast Region
Data extracted 7/27/07
Enrollment
Excluded from Random Assignment (N=6271)
Pre-assignment screening on 8/27/07
• Invalid Police Photo Number/Unique Identifier (n= 431)
o ZZ – 244
o All 9's – 170
o Not found in CPCMS – 8
o No criminal history in CPCMS – 9
• Risk reliability score exclusions (n= 4540)
o Missing reliability score – 576
o Score is less than or equal to 50% – 3964
• Assigned to existing low risk caseload (n=177)
• Offenders with both low and non-low risk cases (n=95)
• Existing potential direct violation (n=320)
• Less than 30 days under supervision at RCT start (n=565)
• Transferred to specialized unit as of 8/27/08 (n=143)
Random Assignment
9/28/07
Experimental Group (n=800)
Transferred to Low Risk Caseload on 10/3/07
Received allocated intervention (n=658)
Did not receive allocated intervention (n=142)
• Non-regional/Specialized unit – 4
• No open cases – 10
• Absconder warrant or 90+ days no contact with officer – 47
• Spanish-speaking only & assigned to Spanish-speaking officer – 2
• FIR (drug evaluation/treatment) condition – 58
• Potential Direct Violation – 16
• Case transferred to another County – 2
• Offender deceased (natural causes)– 2
• Held back from transfer by Director due to poor casework on part of previously assigned officer – 1
Control Group (n= 759)
Received allocated intervention (n=639)
Did not receive allocated intervention (N=120)
• Non-regional/Specialized unit – 8
• No open cases – 10
• Absconder warrant or 90+ days no contact with officer – 40
• Spanish-speaking only & assigned to Spanish-speaking officer – 2
• FIR (drug evaluation/treatment) condition – 50
• Potential direct violation – 10
Lost to follow-up (n=0)
Discontinued intervention (n=0)
Lost to follow-up (n=0)
Discontinued intervention (n=0)
Analyzed (n=800)
Excluded from analysis (n=0)
Analyzed (n=759)
Excluded from analysis (n=0)

FIGURE 20.1. CONSORT diagram of the Philadelphia low-intensity probation RCT (source: Barnes et al. 2010).

The file-drawer problem is a major obstacle to systematic reviews of the literature that are essential to science (see section "Synthesizing Results"). If for no other reason than to reduce the potential bias in reporting the results of research, a registry of randomized trials is needed in criminology to encourage all tests to be made transparent by design and not by results. Once such a registry (or registries) exists, then criminology journals can do what medical journals have recently done. They can require that no publication will be allowed unless a trial was registered publicly in advance of its implementation.

* * * * * * * * * *

This section closes our discussion of the "front end" of experimental criminology. The remainder of this chapter is a brief overview of each of the key issues and arguments that a protocol can anticipate and plan for. Anyone using our Crim-PORT in Appendix 1 may find the next 12 sections especially useful.

"CONTRACTS": RECRUITING, CONSULTING, AND TRAINING KEY PEOPLE

Using the construction metaphor, experiments must be built by contractors (who may also have been the architect) who know how to "contract" with the right people (for money or *pro bono* love of the research). The person in the metaphorical role of building contractor in an experimental criminology is the principal investigator.

Principal investigators in experimental criminology will generally be PhD-level academics. While Jonathan Shepherd and others have espoused a practitioner–investigator model as a means of building a research culture in criminal justice agencies, the current structure of these agencies militates against that model. Anyone in a position to launch a controlled experiment in, for example, a police agency is unlikely to have the time to design an experimental protocol. Nor are they likely to benefit as much in their police careers from publishing a research project as any academic would – unlike doctors in academic medicine. Thus we write the Crim-PORT with an academic experimenter in mind, even as we look to the day when experimenters will be found within crime prevention and justice agencies.

The experimental criminologist – or "random assigner," as Carol Weiss ironically calls all experimentalists – serving as principal investigator should be supported and held accountable as the primary leader of the experiment. While in some experiments an independent evaluator is assigned to measure the results of the experiments (e.g., Shapland et al. 2008), even then the directors of the RCTs must take responsibility for making things happen. When there is no person clearly in that role, the chances of a major failure may increase. In one police experiment, for example, the data were all to be analyzed by a researcher who rarely visited the project site. The research coordinator on the ground was not a PhD and had little status as a temporary employee of the police department. More important, the coordinator had not been hired by the data analyst and had no prior relationship. During the experiment, independent audits of the data showed that there was systematic misrecording of results and ways by which the officers could learn of random assignment in advance of declaring cases eligible.

As this case illustrates, experiments cannot just be "put in the oven and removed when fully baked." They must be steered every moment like an automobile. Alternatively, they must be managed like a construction site. Academics who work at arms length from the randomly assigned operations cannot know what problems exist in the field or how to interpret the data. They need to have an independent source of intelligence on the ground.

Field Coordinators

A field coordinator or manager is the eyes and ears of a distant academic. Only by constant communication with a field coordinator can the principal investigator be sure that the right decisions are taken as crises or routine problems arise. Just as an architect needs to have a good working relationship with a contractor, a principal investigator needs a high level of skill – and trust in – the field coordinator.

This principle is so important that a major foundation recently insisted on giving a leading academic more money than requested to run an experiment at a great distance. The foundation insisted that the academic either hire a competent field coordinator to work daily within the experimenting agency or receive no grant at all. The academic accepted the extra funding, and the experiment was a success.

Other Key Roles

The field coordinator can often be hired with local knowledge about key actors. This helps the academic investigator to do a better job at recruiting key staff and partners. The field coordinator must also be at the center of all social networks connected to the experiment, such as the *agency liaison* and any oversight groups. Some may have a PhD, some may be working on a PhD, and some may not aspire to a PhD. But all should be emotionally and politically intelligent, pragmatic problem solvers of the highest integrity. The same can be said for the *data manager*, who is best kept off-site from the experiment. By reporting independently to the principal investigator, the data manager becomes a second source of check and balance on the experiment. Contacts with the *agents supplying the cases*, in particular, can provide important clues to the progress of the experiment.

Even before recruiting key agency and research people for an experiment, a principal investigator can consult widely among local VIPs to seek advice about who to recruit for key roles. Often the agency liaison can recommend a local research coordinator. Even if advertising is the only source of candidates, enlisting the agency liaison to help choose a field coordinator can help build commitment to the experiment. The same is true for selecting operating agents to apply the treatment to the extent that they can be limited to volunteers. In one relatively small experiment, Minneapolis Police Chief Anthony Bouza let Lawrence Sherman (as principal investigator) choose the entire 5-police officer team for the 2-year experiment. The team included the sergeant in command, who became the primary agency liaison. The team then worked closely with the (local) research coordinator (Dr. Michael Buerger) and the (distant) data manager (Dr. Patrick Gartin), both of whom were doctoral students at the time. All showed sustained commitment to the experiment, from start to finish.

BUILDING: STARTING AND SUSTAINING THE EXPERIMENT

How an experiment begins depends on how it must be designed. There is nothing easier than launching a "batch" random assignment project and nothing harder than launching a "trickle-flow" random assignment project. While months or years of preparation may be required for batch random assignment, they sometimes offer a capacity to literally push a button that will launch the experiment. Random assignment of thousands of letters to taxpayers is a prime example of the batch model.

Trickle-flow experiments, in contrast, may require the cooperation of hundreds of people to supply eligible cases for random assignment. It may be essential to reach them with as much publicity as possible. A press conference, assembly in-person of large groups of agency staff, letters sent to all agents, a launch party with good food and drink, an overnight planning conference at a resort venue, and many other methods have all been used with varying degrees of success. Anything possible that alerts people with access to the raw materials of the experiment is worth trying to encourage them to make "donations" of cases to the cause.

Once the experiment is launched, the even greater challenge with trickle-flow experiments is to sustain the caseflow. The key method is to keep reminding agents who can supply cases to do so. In Sherman's Minneapolis and Milwauke domestic violence experiments, monthly meetings with the research team kept caseflow moving at a steady pace. In Strang and Sherman's Canberra restorative justice experiments, the local police would not allow us to hold meetings with agents who could supply cases. We could only use informal contacts in police agencies where restorative justice meetings were held to remind officers to call our 24-h phone line to refer any eligible cases they encountered. Despite a steady decline in the rate of case referrals, we nonetheless managed to recruit some 1,400 cases over 5 years.

Even batch designs, however, face challenges in launching and sustaining the experiments. In the Minneapolis hot spots experiment (Sherman and Weisburd 1995), the research team spent months checking the units of analysis for eligibility (Buerger et al. 1995). The result was an extremely clean set of visually independent units of analysis. The major challenge was in sustaining a consistent difference in the delivery of the patrol treatment between experimental and control locations over an entire year. The trickle flow in this case was not of new *cases* into the experiment, but of daily treatment dosage into the units of analysis.

Even the selection of batch assignment cases can be problematic if not all the information is available on record. In the Philadelphia Low-Intensity Probation Experiment (Barnes et al. 2010), there was a substantial rate of error in the data entry on characteristics that were keys to the eligibility of probationers for the experiment. Even though the experimenters who checked the eligibility requirements followed all rules precisely, the errors in the data they examined were not made apparent until after random assignment, when treatment modifications were attempted.

Finally, the Philadelphia experiment also illustrates a repeated batch design. This procedure is necessary if the cases age out of the treatment condition as their probation sentences expired, but a certain ratio of offenders to probation officers was part of the definition of the experimental treatment. In order to keep the ratio constant over time, new cases that were not included in the experimental analysis had to be assigned each month or so to the probation officers.

SUPPLYING: OBTAINING CASES

The biggest challenge in trickle-flow experiments is to find and extract the cases that are "leaking" out of the experimental sample. Increasingly, information technology applications can be used to identify the missed cases and the agents who could have referred them to the experimenters. If the agency strongly supports the experiment, there may be ways to encourage agents who do not refer the cases to start doing so. But if the agency will not, or cannot, attempt to persuade those who can contribute cases to an experiment, the only tool left is the ingenuity of the site coordinator or principal investigator.

In Northumbria, for example, neither the Crown Prosecution Service nor the Court Service referred cases to a national government experiment, despite promising to do so. But Dr. Newbury-Birch, the site coordinator, negotiated with the Probation service to have them fax to our research office a copy of any request from the court for a presentence report. That allowed us to directly approach the offenders who had pled guilty in order to seek their consent to be randomly assigned to a restorative justice conference.

SCREENING FOR ELIGIBILITY

"If you break it, you own it," as Colin Powell famously said about the US-led invasion of Iraq in 2003. The same can be said for random assignment of cases that were not eligible for the experiment. Once a case is randomly assigned within a sequence of random numbers, it cannot be deleted without introducing selection bias – exactly the kind of bias random assignment is designed to control.

There is a widespread, mistaken assumption by economists and other analysts that ineligible cases can simply be deleted after random assignment. Samples "corrected" in this way cannot then be analyzed as if the random assignment had never occurred. The only honest correction is to specify a multivariate model that considers both intention-to-treat (ITT) and treatment received (TR), along with any variables that might predict a gap between those two conditions (Piantadosi 1997: 276–282). Taking this path, however, entails the same requirement to be smart and lucky, that is the reason to conduct a randomized trial. Despite advances in statistical thinking about what is called "instrumental variables" to help strengthen causal inference (Angrist 2006), sample sizes large enough to use this approach are in short supply in criminology.

The best solution to the problem of including ineligible cases is to prevent it prior to random assignment. Like so many problems in randomized field experiments, this one can be minimized by better planning and protocols. Hence the best time to exclude ineligible cases is when writing the *budget*: making sure that you spend enough money to have an independent check on the eligibility of cases.

In our Canberra restorative justice experiments (Sherman et al. 2000), the budget included the research team staffing a 24 h a day, 365 days a year telephone hotline for police to call in eligible cases. The first thing staff members did when they answered the phone was to run through a checklist of eligibility requirements. While this rarely detected ineligible cases, it did detect some that were never randomly assigned. The only ineligible cases that slipped through the screening process were those in which police provided inaccurate data: offender's date of birth or other pending charges, for example. The budget for the eligibility screening was also justified by an even more important element of protocol planning: an independent process of random assignment.

RANDOMLY ASSIGNING TREATMENTS

Sherman's first randomized experiment (Sherman and Berk 1984) produced a major finding: saving money on random assignment costs is penny wise and pound foolish. Sherman designed the experiment so that all random assignment sequences were given in advance, without allocation concealment, to each of the agents referring cases and delivering treatments. In that experiment, they were the same people, all Minneapolis patrol officers. This procedure saved money on staffing costs to answer the telephone. But it also allowed officers to selectively include cases when they knew they had an arrest up next in their sequence.

As Gartin (1992) has shown, at least some of the experimental officers in Minneapolis apparently held on to their arrest cases until they encountered a suspect they did not like – or perhaps already knew. In any case, the cases "randomly" assigned to arrest, in proper sequence, for some reason had higher levels of prior arrests than the offenders randomly assigned to nonarrest alternatives. This fact did not alter the substantive conclusion that arrest deterred repeat offending relative to other treatments. Prior arrests predict a higher likelihood of recidivism. But the group that had lower recidivism had more prior arrests. This means that all the officers did was to cause an underestimate of the deterrent effect of arrest in Minneapolis, not an overestimate.

This question never arose in Milwaukee, in which Sherman had a larger budget for random assignment. The chance for biased selection of eligible cases was virtually eliminated by having research staff answer the phone by a secure computer, take the identifying details of the officers and the suspects, and then open a numbered envelope sealed with red sealing wax. Close supervision each day ensured that no envelopes were opened in advance. Analytic checks for prior record and other differences across treatment groups found no significant variation in proportions of such baseline characteristics.

Yet even after Minneapolis, other arrest experiments saved money on random assignment costs by delegating the job to police dispatchers. The theory was that the call record would create transparency about what was known prior to random assignment. What a dispatch center call record would not reveal, however, was the cases that were left out if the dispatcher told the police what assignment was next on an open list. One experiment solved this problem by having a computerized random assignment program programmed right into the dispatch system, so that treatment instructions would not be generated until the identifying case details were registered. Even then, the credibility of an independent random assignment system is well worth the increased budget. A complete firewall of social relations between the operating agents applying treatments and the staff applying random assignments is a policy above suspicion. In our UK restorative justice experiments (Sherman et al. 2005), we even put the random assignment computer and staff on another continent.

Far worse systems of random assignment have been employed. Tossing coins, using dates of birth or days of the week (odd or even dates) – these are all systems that appear to have sufficient integrity to create equal probability of assignment. The general experience is that such systems, like Sherman's in Minneapolis, are more likely to produce differences in case characteristics between randomly assigned groups. Such differences can always happen by chance, of course, especially in a relatively smaller sample. But they are easily prevented by designing a protocol that separates random assignment from operating staff.

DELIVERING TREATMENTS CONSISTENTLY

A great body of literature discusses the role of heterogeneity in the statistical power of experiments. The more differences within groups, the less power there is in the test (Weisburd 1993). This applies to both cases and treatments. Anything that creates differences within the units of analysis, or in the way treatments are delivered, can cause a misleading result: no significant difference despite a "true," underlying difference. While eligibility criteria can limit the differences in cases (for example, in age or prior record), it is much harder for a protocol to insure consistency in the treatment.

The importance of this issue depends in part on the subtlety of the treatment being tested. Something quite blunt, like a decision to arrest, may be full of subtlety (see Paternoster et al. 1997). But the major feature of treatment is taking someone to jail. That is a transparent and easily auditable feature of treatment. Analysts can readily compare cases in which arrest did or did not happen.

In longer and complex treatments, auditing is much harder to attain. Treatments requiring repeated contact can count the number of contacts attained. But they are unlikely to be able to audit the number of minutes the contact lasted, what was discussed, whether participants cried or got angry, or whether their feelings toward treatment staff improved or worsened (Sherman and Strang 2004). All of these things may be important theoretically. If they are – as they were in our Canberra restorative justice experiments – the best plan is to invest heavily in measurements, such as observations and interviews. Even then, however, it is much harder to *deliver* consistency than to *measure* consistency.

What we learned by repeating our Canberra experiments in England was that the protocol mattered. In Canberra, the first police chief was committed to a "generalist" view of policing, in which restorative justice is a skill every police officer should have. At his request, we saw that over 400 officers were trained in the method, and most of the treatments were delivered by officers who delivered them only once. By contrast, the English experiments provided enough funding to have fulltime specialists delivering the same treatment. Even though the same trainers were used to train restorative justice staff in both Canberra and England, the consistency was far higher in England. What created consistency was a specialized unit of "professional" restorative justice facilitators, all of whom having extensive practice. In contrast, delivery of a complex treatment by a "volunteer" created well-documented inconsistency.

MEASURING TREATMENTS DELIVERED

The failure to measure treatment delivery is one of the most common in experimental criminology. Numerous experiments assume that once treatment is assigned it will be delivered. Yet when budgets are invested in measuring delivery, the research shows at least some portion of cases in which delivery did not occur.

The rapid development of information technology will help to lower the costs of treatment measurement. In the Minneapolis Hot Spots patrol experiment, Sherman and Weisburd (1995) invested large sums of National Institute of Justice funds in trained observers with stop watches observing street corners. Their job was to count the number of minutes that police officers were present at each street corner, with arrival and departure times for each "presence," as well as to count crimes and disorders observed.

As we now redesign the experiment with police in Greater Manchester (England), we can save great sums by using two kinds of electronic data. One is the Automatic Radio Locator System (ARLS) that will record where each and every police officer is at all times. This will produce exact counts of minutes that their radio (which they keep on their person at all times for safety reasons) is located at each point in the jurisdiction, by GPS (Global Positioning Satellite) transmissions. The other technology is CCTV cameras, which are trained on many hot spots and can record what happens 100% of the time. High-speed coding software can review the videos of human behavior in high-crime hot spots, thereby saving even data entry costs.

Police presence, however, is a fairly blunt treatment. More subtle treatments will be harder to measure with electronic technologies. The Police Foundation's experiment in counseling officers who were frequently subject to complaints (Pate et al. 1976) entailed discussions behind closed doors with older officers who were allegedly "reformed" from their "cowboy" days on the street. The discussions were supposed to encourage younger officers to keep their tempers and to let insults pass without response. When the experiment backfired – with "counseled" officers suffering more rather than fewer complaints than uncounseled officers – the explanation could not be extracted from the data. The reason was because there were no data. The counseling officers had insisted on no independent observations of the counseling sessions. The result was a classic black box. And the value of the experiment from the standpoint of police service delivery was nil.

The main conclusion that can be derived from such experiments is that investment in treatment measurement is well worth the cost. Without knowing what the treatment truly comprises, there is no way to build upon the results of the experiment. One word for this idea is "descriptive validity," a useful concept for considering the budget issues in every protocol.

MEASURING OUTCOMES

Experimental criminology is blessed with a plethora of official records about crime, as well as growing emphasis on medical records (from emergency rooms), victim interviews, offender self-reports, and observational measures such as CCTV. The challenge is to make sense of them. Several principles may help.

Choose Universal Measures Over Low Response Rate Measures

In studies with two competing measures, findings of each are often given equal weight. This equality may be fine if both are universal measures, such as hospital records and arrest records. They are universal because everyone in the jurisdiction is subject to the record keeping, whatever limitations that may entail. This is true both before and after random assignment, which is unlikely to affect the data collection. But if one measure depends on interviews, it cannot be universal. All interviews are subject to sampling biases from nonresponse. And unlike the universal measures, these biases may be reactive to randomly assigned differences in treatments.

In our Canberra experiments, we had differential response rates from interviews of offenders receiving different treatments. We learned much about offender perceptions of treatments by conducting the interviews, despite the response rate issues. But as measure of repeat offending, we relied more heavily on the universal measures. Whatever biases they suffered, there was far less suggestion of reactivity to treatment.

Choose Crime Frequency Over Prevalence

Crime prevention experiments are ultimately aimed to reduce the crime problem in communities. When individual offenders are the unit of analysis, the best measure of their effect on community crime is the frequency and seriousness of their offending. Yet by tradition, many government agencies are locked into a precomputer definition of "recidivism" as the *prevalence* of repeat offending: what percentage of offenders in each group had one or more arrest or conviction during the follow-up period. We know many instances in which the findings showed no differences in prevalence, but with large differences in frequency. The UK Home Office only used prevalence criteria for many years to judge any program a success. Only when Shapland and her colleagues (2008) discovered that our restorative justice experiments had reduced the frequency of crime (but not its prevalence) did the policy of exclusive focus on prevalence change. As an indicator of how much crime occurs in the community in response to treatments, frequency seems to be the far more sensitive and reliable indicator.

Choose a Seriousness Index Over Categorical Counts

The larger problem with measuring community benefit in criminology is the general failure to weight the seriousness of crimes. Crimes vary widely in their costs or perceived seriousness. On this measure, in fact, we have had our greatest success, with the cost effectiveness of our UK experiments showing a significant 9 to 1 return on investment (Shapland et al. 2008). But this too has not been a traditional measure of success, even though the UK has some of the best-developed measures of the average costs of crime of any country.

Choose One Measure as Primary at the Outset

The best way to avoid arguments at the end of an experiment is to agree at the outset what the primary outcome measure will be. Had we agreed on cost effectiveness based on seriousness and frequency of crime as the key criterion for a US $10 million set of experiments, the public would have been better served in its investment in testing restorative justice. But even this is an elusive goal, since funding personnel changed so often in the course of the 5-year project. What may matter most in the long run is a development of consensus within professional groups, such as the Campbell Collaboration. Financial agencies of governments, such as the US Office of Management and Budget or the Treasury in the UK, could also support a greater focus on cost effectiveness rather than the kind of bean-counting of prevalence (or even frequency) that Sherman and other experimentalists have done in the past.

Perhaps the best reason to pick a primary measure in advance is the frequent debate about whether data analysts "fished" for a significant result, highlighting one significant difference in a long procession of null findings. Gorman and Huber (2009) have recently demonstrated that even a program widely believed (on the evidence) to be ineffective (DARE) can be shown to be effective by the same analytic methods used to report programs widely believed to "work." Advance registration of a protocol following the Crim-PORT would prevent this problem before an experiment even begins.

ANALYZING RESULTS

Many statistics textbooks address this question. We have two comments to supplement those texts. The first is to seek simplicity in analysis. Experiments are elegant in their simplicity. They can also be analyzed that way. The simpler the analysis, the more people will be able to grasp the meaning of the results. Cost effectiveness is attractive on these grounds as well as on the substantive grounds of public benefit. Complex statistical models and uninterpretable effect sizes are not.

The second comment is to test policies, not treatments. This means, in general, that Intention-To-Treat (ITT) analysis makes the most sense in keeping the analysis simple. Experiments in criminology will always feature complexity of how people deliver treatments and react to them. But testing a decision to follow one *policy* to attempt one treatment or another with each case does not require that the treatment actually occurs. What happens after the attempt-to-treat begins is actually an answer to the question posed by the experiment. That question can be answered just fine by ITT analysis. So can the question about what effect the policy had on outcomes. What the experiment cannot answer – absent near-perfect delivery of the treatment – is what effect the treatment had.

Why is policy more important than treatment? In the long run, treatment delivery could be improved and policies of offering it could have very different effects. If that happened, then new experiments would be needed to test the effect of the treatment-enriched policies. Thus as long as the experiment is limited to the random assignment of policy and cannot control treatment, the honest thing to do is to analyze the effects of policy.

COMMUNICATING RESULTS

Simplicity is also a great virtue in communicating results. Academics inclined to making fine distinctions are often impatient with simplicity. But they lack evidence to support the claim that complexity will lead to better policymaking or even better science. We know few academics who can recall the details of any particular study unless it directly pertains to their own research of the moment. The error rate in print in describing our own research is also very high, even among distinguished scholars. Those who attack simple conclusions, like saying something "works" or not, may be praised for their lofty aspirations for the future of human intelligence.

SYNTHESIZING RESULTS

The best reason for doing experiments is that they have value far beyond the era in which they are completed and reported. This value is sometimes limited to the single experiment. More often, however, the value of each experiment grows as replications and related research accumulates (Sherman and Strang 2004a). And as experimental criminology has grown, so too has the related field of research synthesis.

The goal of research synthesis is to draw conclusions from a universe of all tests of a single hypothesis. The Campbell Collaboration is promoting this task in crime and justice, as well as in other social policy areas. Randomized experiments are especially valuable for systematic reviews and meta-analysis of accumulated tests of a program or policy. Even those who are generally critical of the statistical basis of meta-analysis are ready to endorse its use when only randomized experiments are included in the calculations (Berk 2005).

There is all the more reason, then, to plan experiments well, to minimize alternative explanations, and to report all the items needed for others to include your results in research syntheses. The experimenter's prime audience is no longer the scholars and leaders of the day. A much larger audience will use solid research results for many decades to come.

BECOMING A RANDOM ASSIGNER

We conclude our introduction to experimental criminology with a recruitment poster. Do you have what it takes to become what Weiss (2002) calls a "random assigner"? Can you tell by now what personal qualities are needed to do this well? Can you tell how the personal experience of experimental criminologists differs from the daily life of other field researchers and from people who analyze existing data?

In our view, experimental criminology requires a more extroverted personality than is needed for scholarship in general. Like experimental physics, EC needs large teams of people to cooperate. Leadership skills are essential to fostering that cooperation. Someone who would enjoy being a university department chair would probably enjoy and do well at EC. But such people are generally rare in academic life.

Experimental criminology may also require a greater readiness to accept the big problems that cannot be changed quickly, in order to attack smaller problems that can be. Social and economic injustice has deep roots and structural support. But much of what we do about it may only make things worse. Experimentalists can at least find better ways to do no harm and perhaps more good.

The best experimental criminology will feature the best traits of scholarship in general: erudition, broad theoretical vision, a nuanced grasp of causal inference, and abiding curiosity. The "sacred spark" that drives all scholarship is especially needed to persevere in the face of the many setbacks that EC suffers. Even Einstein might see it as a field involving more than "mere experimental skill."

So think it over. Experimental criminology is looking for a few good women.

And men.

Further Readings: Boruch, 1997; Gottfredson et al., 2006; Sherman, 2009.

REFERENCES

Angrist JD (2006) Instrumental variables methods in experimental criminological research: What, why and how. J Exp Criminol 2(1):23–44

Angrist J, Imbens G, Rubin D (1996). J Am Stat Assoc 91:444 – 455

Ares CE, Rankin A, Sturz H (1963) The Manhattan Bail Project: an interim report on the use of pre-trial parole. N Y Univ Law Rev 38:67–95

Ariel B (2008) Seminar presented to the Jerry Lee Centre for Experimental Criminology, Institute of Criminology, University of Cambridge, October

Ariel B (2009) Taxation and compliance: an experimental study. Doctoral Dissertation, Hebrew University of Jerusalem, Israel

Barnes G, Ahlman L, Gill C, Kurtz E, Sherman L, Malvestuto R (2010) Low-intensity community supervision for low-risk offenders: a randomized, controlled trial. Journal of Experimental Criminology, forthcoming

Berk RA (2005) Randomized experiments as the bronze standard. J Exp Criminol 1(4):417–433

Bliss M (1999) William Osler: a life in medicine. University of Toronto Press, Toronto

Boruch RF (1997) Randomized experiments for policy and planning. Sage, Newbury Park, CA

Bradley RS, Jones PD (1992) Climate since A.D. 1500. Routledge, London

Braithwaite J (1989) Crime, Shame and Reintegration. Cambridge: Cambridge University Press

Braithwaite J (2002) Restorative Justice and Responsive Regulation. NY: Oxford U. Press

Buerger M, Cohn E, Petrosino A (1995) Defining the "hotspots of crime": operationalizing theoretical concepts for field research. In: Eck JE, Weisburd D (eds) Crime and place. Crime Prevention Studies, vol 4. Police Executive Research Forum. Criminal Justice Press, Monsey, NY

Campbell DT, Stanley JC (1963) Experimental and quasi-experimental designs for research. Rand-McNally, Chicago, IL

Cook TD, Campbell DT (1979) Quasi-experimentation: design and analysis issues for field settings. Rand-McNally, Chicago

Einstein A, Infeld L ([1938] 1971) The evolution of physics, 2nd edn. Downloaded at Google Books on 12 July 2009

Eisner MP (2009) No effects in independent prevention trials: can we reject the cynical view? J Exp Criminol 5(2):163–183

Erwin BS (1986) Turning up the heat on probationers in Georgia. Fed Probat 50:17–24

Farrington DP (1983) Randomized experiments on crime and justice. In: Tonry M, Morris N (eds) Crime and justice: an annual review of research, vol 4. University of Chicago Press, Chicago, IL

Farrington DP (2003) British randomized experiments on crime and justice. Ann Am Acad Polit Soc Sci 589:150–169

Farrington DP, Knight BJ (1980) Stealing from a "Lost" letter: effects of victim characteristics. Crim Justice Behav 7:423–436

Farrington DP, Welsh BC (2005) Randomized experiments in criminology: what have we learned in the last two decades? J Exp Criminol 1(1):9–38

Federal Judicial Center (1981) Experimentation in the law. Federal Judicial Center, Administrative Office of the US Courts, Washington, DC

Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh

Gartin PR (1992) A Replication and Extension of the Minneapolis Domestic Violence Experiment. PhD. Dissertation, University of Maryland

Gibbs JP (1975) Crime, punishment and deterrence. Elsevier, New York

Gladwell M (2005) Blink: the power of thinking without thinking. Little, Brown, Boston, MA

Gladwell M (2008) Outliers: the story of success. Little, Brown, Boston, MA

Glaser BG, Strauss AL (1967) The discovery of grounded theory: strategies for qualitative research. Aldine Publishing Company, Chicago, IL

Gorman DM, Huber JC (2009) The social construction of "evidence-based" drug prevention programs: a reanalysis of data from the Drug Abuse Resistance Education (DARE) Program. Eval Rev 33:396–414

Hanley D (2006) Appropriate services: examining the case classification principle. J Offender Rehabil 42:1–22

Home Office (2005) The economic and social costs of crime against individuals and households 2003/04. Home office on-line report 30/05 downloaded on 26 July, 2009 from http://www.homeoffice.gov.uk/rds/pdfs05/rdsolr3005.pdf

Kirk DS (2009) A natural experiment on residential change and recidivism: lessons from hurricane Katrina. Am Sociol Rev 74(3):484–504

Laub J, Sampson R (2003) Shared Beginnings, Duivergent Lives. Cambridge: Harvard University Press

Loudon I (2002) Ignaz Phillip Semmelweis' studies of death in childbirth. The James Lind Library (http://www.jameslindlibrary.org). Accessed FxTuesday 4 August 2009

Palmer T, Petrosino A (2003) The "experimenting agency". The California Youth Authority Research Division. Eval Rev 27:228–266

Pate T, McCullough JW, Bowers R, Ferra A (1976) Kansas City Peer Review Panel: An Evaluation Report. Washington, DC: Police Foundation

Paternoster R, Brame R, Bachman R, Sherman L. (1997) Do fair procedures matter? The effect of procedural justice on spouse assault. Law Soc Rev 31(1):163–204

Piantadosi S (1997) Clinical trials: a methodologic perspective. Wiley, New York

Pocock SJ, Hughes MD, Lee RJ (1987) Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 317:426–432

Salsburg D (2001) The lady tasting tea: how statistics revolutionized science in the twentieth century. Henry Holt, New York

Shapland J, Atkinson A, Atkinson H, Dignan J, Edwards L, Hibbert J, Howes M, Johnstone J, Robinson G, Sorsby A (2008) Does restorative justice affect reconviction? The fourth report from the evaluation of three schemes. Ministry of Justice Research Series 10/08, June. Ministry of Justice, London

Sherman LW (1979) The case for the research police department. Police Mag 2(6):58–59

Sherman LW (1992) Policing Domestic Violence: Experiments and Dilemmas. NY: Free Press

Sherman LW (1993) Defiance, deterrence and irrelevance: a theory of the criminal sanction. J Res Crim and Delin 30:445–473

Sherman LW (2006) To develop and test: the inventive difference between evaluation and experimentation. J Exp Criminol 2:393–406

Sherman LW (2007) The power few: experimental criminology and the reduction of harm. The 2006 Joan McCord Prize Lecture. J Exp Criminol 3(4):299–321

Sherman LW (2009) Evidence and liberty: the promise of experimental criminology. Criminol Crim Justice 9:5–28

Sherman LW, Berk RA (1984) The specific deterrent effects of arrest for domestic assault. Am Sociol Rev 49:261–271

Sherman LW, Rogan DP (1995a) Effects of gun seizures on gun violence: "hot spots" patrol in Kansas city. Justice Q 12(4):673–693

Sherman LW, Rogan DP (1995b) Deterrent effects of police raids on crack houses: a randomized controlled experiment. Justice Q 12(4):755–781

Sherman LW, Strang H (2004a) Verdicts or inventions? Interpreting randomized controlled trials in criminology. Am Behav Sci 47(5):575–607

Sherman LW, Strang H (2004b) Experimental ethnography: the marriage of qualitative and quantitative research. In: Anderson E, Brooks SN, Gunn R, Jones N (eds) Annals of the American academy of political and social science, vol 595, pp 204–222

Sherman LW, Strang H (2010) Doing experimental criminology. In: Gadd D, Karstedt S, Messner S (eds) Handbook of criminological research methods. Sage, Thousand Oaks, CA

Sherman LW, Weisburd D (1995) General deterrent effects of police patrol in crime hot spots: a randomized, controlled trial. Justice Q 12(4):635–648

Sherman LW, Smith DA, Schmidt J, Rogan DP (1992) Crime, punishment and stake in conformity: legal and informal control of domestic violence. Am Sociol Rev 57:680–690

Sherman LW, Strang H, Woods D (2000) Recidivism patterns in the Canberra reintegrative shaming experiments (RISE). Downloaded on 5 August 2009 at http://www.aic.gov.au/criminal_justice_system/rjustice/rise/aspx

Sherman LW, Strang H, Angel C, Woods D, Barnes G, Bennett S, Rossner M, Inkpen N (2005) Effects of face-to-face restorative justice on victims of crime in four randomized controlled trials. J Exp Criminol 1(3):367–395

The Multisite Violence Prevention Project (2008) Impact of a universal school-based violence prevention program on social-cognitive outcomes. Prev Sci 9(4):231–244

Tilley N (2009) Sherman vs Sherman: realism vs rhetoric. Criminol Crim Justice 9(2):135–144

Tröhler U (2003) 'James Lind and Scurvy: 1747 to 1795.' The James Lind Library (http://www.jameslindlibrary.org). Downloaded 4 August, 2009

Weisburd D (1993) Design sensitivity in criminal justice experiments. Crime and Justice 17:337–379

Weiss C (2002) What to do until the random assigner comes. In: Mosteller F, Boruch R (eds) Evidence matters. Brookings Institution, Washington, DC

Welsh BC, Farrington DP, Sherman LW (2001) Costs and benefits of preventing crime. Westview Press, Boulder, CO

Appendix 1

CRIM-PORT 1.0:

Criminological Protocol for Operating Randomized Trials

INSTRUCTIONS: Please use this form to enter information directly into the WORD document as the protocol for your registration on Cambridge University's Jerry Lee Centre of Experimental Criminology's *Registry of EXperiments in Policing Strategy and Tactics* (REX-POST) or the separate *Registry of Experiments in Corrections Strategy and Tactics* (REX-COST) at http://www.crim.cam.ac.uk/experiments.

CONTENTS:

6. Timing
7. Random Assignment
8. Treatment and Comparison Elements
9. Measuring and Managing Treatments
10. Measuring Outcomes
11. Analysis Plan
12. Due Date and Dissemination Plan

1. Name and Hypotheses

A. **Name of Experiment**________________________________
B. Principal Investigator
(Name)____________________________
(Employer)________________________________
C. 1st Co-Principal Investigator
(Name)____________________________
(Employer)________________________________
D. 2d Co-Principal Investigator (Name)____________________________
(Employer)________________________________
E. **General Hypothesis**: (Experimental or Primary Treatment) _________ causes (less or more) ______ (crime or justice outcome) _________ than (comparison or control treatment) _________.
F. **Specific Hypotheses**:

1. List all variations of treatment delivery to be tested.
2. List all variations of outcome measures to be tested.
3. List all subgroups to be tested for all varieties of outcome measures.

2. Organizational Framework: Check only one from a, b, c, or d

A. **In-House** delivery of treatments, data collection and analysis ___
B. **Dual Partnership**: Operating agency delivers treatments with independent research organization providing random assignment, data collection, analysis ___

Name of Operating Agency____________________________
Name of Research Organization____________________________

C. **Multi-Agency Partnership**: Operating agencies delivers treatments with independent research organization providing random assignment, data collection, analysis

Name of Operating Agency
1____________________________
Name of Operating Agency
2____________________________
Name of Operating Agency
3____________________________
Name of Research
Organization____________________________

D. **Other Framework** (describe in detail).

3. Unit of Analysis

Check only one

__A. People (describe role: offenders, victims, etc.)________________________
__B. Places (describe category: school, corner, face-block, etc.)____________
__C. Situations (describe: police-citizen encounters, fights, etc.)____________
__D. Other (describe)________________________________

4. Eligibility Criteria

A. **Criteria Required** (list all)
B. **Criteria for Exclusion** (list all)

5. Pipeline: Recruitment or Extraction of Cases (answer all questions)

A. Where will cases come from?
B. Who will obtain them?
C. How will they be identified?
D. How will each case be screened for eligibility?
E. Who will register the case identifiers prior to random assignment?
F. What social relationships must be maintained to keep cases coming?
G. Has a Phase I (no-control, "dry-run") test of the pipeline and treatment process been conducted? If so,

- How many cases were attempted to be treated
- How many treatments were successfully delivered
- How many cases were lost during treatment delivery

6. Timing: Cases come into the experiment in (check only one)

A. A trickle-flow process, one case at a time____
B. A single batch assignment____
C. Repeated batch assignments____
D. Other (describe below)____

7. Random Assignment

A. How is random assignment sequence to be generated?
(coin-toss, every Nth case, and other nonrandom tools are banned from CCR-RCT).
Check one from 1, 2 or 3 below

1. Random numbers table → case number sequence → sealed envelopes with case numbers outside and treatment assignment inside, with 2-sheet paper surrounding treatment___
2. Random numbers case–treatment generator program in secure computer___
3. Other (please describe below)___

B. Who is entitled to issue random assignments of treatments?
Role:
Organization:
C. How will random assignments be recorded in relation to case registration?
Name of data base:
Location of data entry:
Persons performing data entry:

8. Treatment and Comparison Elements

A. Experimental or Primary Treatment

1. What elements must happen, with dosage level (if measured) indicated.
   Element A:
   Element B:
   Element C:
   Other Elements:
2. What elements must *not* happen, with dosage level (if measured) indicated.
   Element A:
   Element B:
   Element C:
   Other Elements:

B. Control or Secondary Comparison Treatment

3. What elements must happen, with dosage level (if measured) indicated.
   Element A:
   Element B:
   Element C:
   Other Elements:
4. What elements must not happen, with dosage level (if measured) indicated.
   Element A:
   Element B:
   Element C:
   Other Elements:

9. Measuring and Managing Treatments

A. Measuring

1. How will treatments be measured?
2. Who will measure them?
3. How will data be collected?
4. How will data be stored?
5. Will data be audited?
6. If audited, who will do it?
7. How will data collection reliability be estimated?
8. Will data collection vary by treatment type?
   If so, how?

B. Managing

1. Who will see the treatment measurement data?
2. How often will treatment measures be circulated to key leaders?
3. If treatment integrity is challenged, whose responsibility is correction?

10. Measuring and Monitoring Outcomes

A. Measuring

1. How will outcomes be measured?
2. Who will measure them?

3. How will data be collected?
4. How will data be stored?
5. Will data be audited?
6. If audited, who will do it?
7. How will data collection reliability be estimated?
8. Will data collection vary by treatment type?
   If so, how?

B. Monitoring

1. How often will outcome data be monitored?
2. Who will see the outcome monitoring data?
3. When will outcome measures be circulated to key leaders?
4. If experiment finds early significant differences, what procedure is to be followed?

11. Analysis Plan

A. Which outcome measure is considered to be the primary indicator of a difference between experimental treatment and comparison group?
B. What is the minimum sample size to be used to analyze outcomes?
C. Will all analyses employ an intention-to-treat framework?
D. What is the threshold below which the percent Treatment-as-Delivered would be so low as to bar any analysis of outcomes?
E. Who will do the data analysis?
F. What statistic will be used to estimate effect size?
G. What statistic will be used to calculate P values?
H. What is the magnitude of effect needed for a $P = 0.05$ difference to have an 80% chance of detection with the projected sample size (optional but recommended calculation of power curve) for the primary outcome measure.

12. Dissemination Plan

A. What is the date by which the project agrees to file its first report on CCR-RCT? (report of delay, preliminary findings, or final result).
B. Does the project agree to file an update every 6 months from date of first report until date of final report?
C. Will preliminary and final results be published, in a 250-word abstract, on CCR-RCT as soon as available?
D. Will CONSORT requirements be met in the final report for the project? (See http://www.consort-statement.org/)
E. What organizations will need to approve the final report? (include any funders or sponsors)
F. Do all organizations involved agree that a final report shall be published after a maximum review period of 6 months from the principal investigator's certification of the report as final?
G. Does principal investigator agree to post any changes in agreements affecting items 12A to 12F above?
H. Does principal investigator agree to file a final report within 2 years of cessation of experimental operations, no matter what happened to the experiment?

(e.g., "random assignment broke down after 3 weeks and the experiment was cancelled" or "only 15 cases were referred in the first 12 months and experiment was suspended").

An Introduction to Experimental Criminology: Lawrence W. Sherman

Background Experimental criminology (EC) is scientific knowledge about crime and justice discovered from random assignment of different conditions in large field tests.

1. This method is the preferred way to estimate the average effects of one variable on another, holding all other variables constant
2. While the experimental method is not intended to answer all research questions in criminology, it can be used far more often than most criminologists assume
   - Opportunities are particularly promising in partnership with criminal justice agencies

Note: The goal of this chapter is to help its readers improve the design and conduct of criminological experiments. This chapter's method is to describe the necessary steps and preferred decisions in planning, conducting, completing, analyzing, reporting, and synthesizing high-quality randomized controlled trials (RCTs) in criminology.

EC use The highest and best use of experimental criminology is to develop and test theoretically coherent ideas about reducing harm (Sherman 2006, 2007), rather than just "evaluating" government programs.

- Those tests, in turn, can help to accumulate an integrated body of grounded theory in which experimental evidence plays a crucial role.
- The advantages depend entirely on the capability of the experimenters to insure success in achieving the many necessary elements of an unbiased comparison:

1. Many randomized field experiments in criminology suffer flaws that could have been avoided with better planning.

Metaphors for experiments The success of experimental criminology may depend on choosing the right metaphor.

- The most useful metaphor is constructing a building.
- The recurrent metaphor of constructing a building helps to illustrate the order of steps to take for best results.
- The steps presented in this chapter begin with the intellectual property of every experiment: formulating the research question.
- Once a protocol is agreed and approved, the experimenters (like builders) must find and "contract" with a wide range of agents and others to best construct and sustain the experiment.
- When and if all these steps are completed, the experiment will be ready for analysis.
- This chapter briefly maps out those principles and the arguments for and against fundamentally different analytic approaches in EC.

Part 1: Intellectual Property: Formulating the Research Question

Great experiments Great experiments in criminology are arguably based on three criteria:

1. They test theoretically central hypotheses – experimentalists can do the most good for science when they are the most focused on the theoretical implications of their experiments.
2. They eliminate as many competing explanations as possible – it is the capacity to limit ambiguity by eliminating competing explanations that makes EC so important to criminological theory.
3. They show one intervention to be far more cost effective than others – rising interest in this principle alone has done more to encourage evidence-based government programs than any other.

   - Experiments must be planned to measure costs of delivering programs, both in a start-up phase and in a "rollout" model with perhaps more efficiencies from mass production.

Note: Putting these criteria together in the formulation of an experimental research question may seem to be more a matter of "art" than of science. Such a judgment would demean the importance of intuition, inspiration, and insight in science, as in many fields involving complex decisions.

Part 2: Social Foundation: Developing a "Field Station"

Field stations The history of experimental field science shows many examples of research centered in what looks much like an indoor laboratory – but with a crucial difference – studies consider questions that cannot be answered in a laboratory

- Field research stations have collected various kinds of observational data systematically in the same places for at least 300 years.

- By the 1950s, hospitals associated with medical schools took on the same character as field stations, linking teaching and research with a large number of clinical randomized controlled trials (RCT)
- From at least the 1960s, similar concentrations of field experiments have been found in the criminal justice system
- The concept of a field station where data are recorded and experiments can last for many decades is an explicit vision for how to conduct experiments in criminology

Social elements The key to holding an experiment together is understanding a cognitive map of its social elements which include the funders, the executive leadership of an operating agency, the mid-level operating liaison person, the agents delivering treatments, and where necessary the agents providing cases.

Part 3: Deciding on the Experimental Protocol

Experiment blueprints The future is clear: experimental criminology will need to design blueprints and stick to them, absent approval from oversight bodies.

- It can be argued that EC will be substantially improved by wider use of experimental protocols.
- One reason is so many RCTs in criminology have either violated good design standards or failed to report fundamental information.
- CONSORT – CONsolidated Standards On Reporting of Trials can lead to better planning of experiments with protocols that anticipate reporting requirements.
- The CONSORT checklist includes 22 reporting elements

Title & Abstract	Background	Participants
Interventions	Objectives	Outcomes
Sample size	Sequence generation	Allocation Concealment
Implementation	Statistical Methods	Participant flow
Recruitment	Baseline data	Numbers analyzed
Outcomes & estimation	Ancillary analyses	Adverse events
Interpretation	Generalizability	Overall evidence
Blinding (Masking)		

Note: CONSORT alone is not enough. A reporting system does not tell you how to design a protocol for an experiment before it starts. It only tells you what readers need to know after it is finished. Included in this chapter is the first version of a standard protocol format for experiments in criminology. The appendix lays out the elements of the protocol.

Part 4: "Contracts": Recruiting, Consulting, and Training of Key People

Contracts Using the construction metaphor, experiments must be built by contractors (who may also have been the architect) who know how to "contract" with the right people (for money or *pro bono* love of the research).

- Principal investigators in experimental criminology will generally be PhD-level academics.
- The experimental criminologist – or "random assigner," serving as principal investigator should be seen supported and held accountable as the primary leader of the experiment
- Other key roles are the field coordinator, the agency liaison, the data manager, and the agent supplying the cases

Part 5: Starting and Sustaining the Experiment

Design How an experiment begins depends on how it must be designed.

- There is nothing easier than launching a "batch" random assignment project, and nothing harder than launching a "trickle-flow" random assignment project.
    1. While months or years of preparation may be required for batch random assignment, they sometimes offer a capacity to literally push a button that will launch the experiment.
    2. Trickle-flow experiments, in contrast, may require the cooperation of hundreds of people to supply eligible cases for random assignment – After launch, the even greater challenge with trickle-flow experiments is to sustain the case flow.

Part 6: Supplying: Obtaining Cases

Supplying cases The biggest challenge in trickle-flow experiments is to find and extract the cases that are "leaking" out of the experimental sample.

- Increasingly, information technology applications can be used to identify the missed cases and the agents who could have referred them to the experimenters.
- If the agency strongly supports the experiment, there may be ways to encourage agents who do not refer the cases to start doing so.
- But if the agency will not, or cannot, attempt to persuade those who can contribute cases to an experiment, the only tool left is the ingenuity of the site coordinator or principal investigator.

Part 7: Screening for Eligibility

Avoiding ineligible cases The best solution to the problem of including ineligible cases is to prevent it prior to random assignment.

- The best time to exclude ineligible cases is when writing the *budget*: making sure that you spend enough money to have an independent check on the eligibility of cases.

Part 8: Assigning Treatments

Assignment system Saving money on random assignment costs is penny wise and pound foolish.

- The chance for biased selection of eligible cases can be virtually eliminated by having research staff answer the phone by a secure computer, take the identifying details of the officers and the suspects, and then open a numbered envelope sealed with red sealing wax
- The credibility of an independent random assignment system is well worth the increased budget.
- It is important to design a protocol that separates random assignment from operating staff.

Part 9: Delivering Treatments Consistently

Protocol matters Anything that creates differences within the units of analysis, or in the way treatments are delivered, can cause a misleading result: no significant difference despite a "true," underlying difference.

- While eligibility criteria can limit the differences in cases (for example, in age or prior record), it is much harder for a protocol to insure consistency in the treatment
- The best plan is to invest heavily in measurements, such as observations and interviews – even then, however, it is much harder to *deliver* consistency than to *measure* consistency
- Repeating experiments done in Canberra Australia and then again in England showed that protocol matters and that as a result, consistency was far higher in England.

Part 10: Measuring Treatments Delivered

Failure to measure The failure to measure treatment delivery is one of the most common in experimental criminology.

- Numerous experiments assume that once treatment is assigned it will be delivered; yet when budgets are invested in measuring delivery, it shows at least some portion of cases in which delivery did not occur.
- Great sums can be saved by using two kinds of electronic data:

  1. One is the Automatic Radio Locator System (ARLS) that will record where each and every police officer is at all times.
  2. The other technology is CCTV cameras, which are trained on many hot spots and can record what happens 100% of the time.

Part 11: Measuring Outcomes

Principles Several principles can help in measuring outcomes.

- Choose universal measures over low response rate measures
- Choose crime frequency over prevalence
- Choose a seriousness index over categorical counts
- Choose one measure as primary at the outset

Part 12: Analyzing Results

Analysis issues There are two issues to supplement what textbooks say on analysis of results.

1. The first is to seek simplicity in analysis.
2. The other is test policies, not treatments – in general, Intention-To-Treat analysis makes the most sense in keeping the analysis simple.

   - As long as the experiment is limited to the random assignment of policy and cannot control treatment, the honest thing to do is to analyze the effects of policy.

Part 13: Communicating Results

Avoid complexity Simplicity is also a great virtue in communicating results.

- Academics inclined to making fine distinctions are often impatient with simplicity, but they lack evidence to support the claim that complexity will lead to better policymaking or even better science.

Part 14: Synthesizing Results

The goal of synthesis The goal of research synthesis is to draw conclusions from a universe of all tests of a single hypothesis.

- Randomized experiments are especially valuable for systematic reviews and meta-analysis of accumulated tests of a program or policy.
- Even those who are generally critical of the statistical basis of meta-analysis are ready to endorse its use when only randomized experiments are included in the calculations

Part 15: Becoming a Random Assigner

Recruitment Who makes the best random assigner?

- Experimental criminology requires a more extroverted personality than is needed for scholarship in general.
- Experimental criminology may also require a greater readiness to accept the big problems that cannot be changed quickly, in order to attack smaller problems that can be.
- The best experimental criminology will feature the best traits of scholarship in general: erudition, broad theoretical vision, a nuanced grasp of causal inference, and abiding curiosity.

Appendix The appendix contains a Criminological Protocol for Operating Randomized Trials.

This information map was prepared with the support of the Jerry Lee Foundation and the assistance of Herbert Fayer.