The Use of Randomized Field Trials in the U.S. to Evaluate Justice Programs: A Briefing

Anthony Petrosino
Research Consultant

Abstract

This article overviews the use of randomized experiments in criminal justice and criminology. The paper presents the rationale for doing experiments, emphasizing its ability to rule out causal explanations for observed findings. After presenting a brief history of justice trials, the paper concludes with a discussion of limitations of the method.

Key words: experimental design, random assignment, program evaluation

Evaluators have developed many helpful approaches to studying the effects of programs, but they may be irrelevant if we lack confidence that an observed impact should be credited to the intervention. In this article, I consider the role of randomized field trials in providing more precise evidence to determine what interventions are effective. Prior criticisms offered by reviewers on the evaluation evidence inspired this paper. I begin with those criticisms and then provide their rationale for recommending randomized trials, with particular attention to the concept of internal validity. After discussing some quasi-experimental approaches to evaluation, the paper highlights the special recognition received by randomized field trials. A brisk history of criminological experiments in the United States follows, after which I discuss some of the research that compares the results of trials to non-randomized evaluations. After summarizing threats to the internal validity of an experiment, I conclude that such controlled studies can address a major problem with original evaluations identified by prior reviewers and other commentators on the criminological evidence.

The critics of evaluation evidence

Policymakers seeking unequivocal information about the effectiveness of programs to reduce crime have been frustrated by the ambiguous and conflicting evidence in the criminological literature. Their conclusions, discouragingly enough, appear similar no matter when they were written. In 1970, the President’s Commission on Law Enforcement and the Administration of Justice reviewed research on the effects of juvenile justice interventions and concluded that “No responsible business concern would operate with as little information regarding its success or failure as do...
nearly all of our delinquency-prevention and control programs" (Wheeler et al. 1970:440). During the 1980s, the MacArthur Foundation funded a group of researchers to examine criminal justice programs and the evidence on their effectiveness. The Justice Study Group, following their review, stated that the knowledge needed to reduce crime “does not exist except in fragmentary and unsatisfactory form...” (Farrington et al 1986:17).

More than ten years later, researchers at the University of Maryland summarized the evidence on crime prevention for the United States Congress (Sherman et al. 1997). Although the University of Maryland report is less pessimistic about the evidence on crime prevention, the lack of rigorous evaluation led the research team to recommend large increases in funding to support stronger outcome studies of federal justice programs (Sherman et al. 1997). The criticisms contained in reports issued by the President’s Commission, the MacArthur Justice Study Group, and the University of Maryland, were not alone. They joined in with the conclusions reached by many other researchers, most of whom had conducted broad surveys and reviews of the treatment literature in criminal justice occasionally over four decades (e.g. Wright and Dixon 1977; Martinson, 1974; Bailey 1966).

The reviews by Martinson (1974) and others are remembered for the skepticism they inspired about the merits of correctional intervention. Each of them concluded that there was little evidence that programs in offender rehabilitation or delinquency prevention worked. Their reviews of the quality of the research evidence, however, were even more critical. For example, Bailey (1966) wrote that the research evidence supporting rehabilitation was “slight, inconsistent, and of questionable reliability” (1966: 157). Martinson’s appraisal of the evaluation evidence was more negative. He wrote that, “it is just possible that our treatment programs are working to some extent, but that our research is so bad it is incapable of telling” (Martinson 1974:49). Wright and Dixon examined 96 evaluations in delinquency prevention and reported that they were of such low quality they caustically concluded that “few of them should have seen the light of day” (1977:57).

**The main problem underscored by critics: low internal validity**

To understand the criticism, we should examine the reviewer’s job. Many of these critics took the task of reading available evaluations and attempting to come to a summary conclusion about that types of interventions seem to be effective in reducing crime. We can sympathize with their plight. No matter what the evaluator of the original program concluded about the effects of the intervention under study, the reviewer often had little confidence that the effect could be attributed to the program. Researchers often refer to confidence in the evaluation’s ability to attribute observed impact to the program as internal validity (Campbell and Stanley 1963). A program evaluation design is often judged on how well it reduces threats - or alternative explanations for observed program impact - to internal validity. Generally, we can be more confident in the conclusions of evaluations with high internal validity than those with low internal validity (Weisburd
and Petrosino, forthcoming)

In other words, the confidence that we place in the evaluation’s internal validity is based on how well the design rules out plausible rival explanations for the results (Campbell and Stanley 1963). A program may be implemented because of prior research findings, the common sense of planners, or the vision of advocates. But without rigorous testing, we are not sure that the program is responsible for changes (if any occur) or that the intervention does not cause harm to its participants (e.g. Finckenauer, 1982).

Universally, reviewers examining study reports wrote that the evaluations were low in internal validity; even large and positive outcomes could not be attributed to the program. For example, Jenson and Howard (1990) reviewed some of the research evidence on social skills training for juvenile delinquents. One study in their review reported large effects but had low internal validity, forcing them to conclude, “...the findings...are subject to many interpretations...” (Jenson and Howard, 1990:223). The reviewer’s job of coming to summary conclusions about what works is hampered in the face of such equivocal evidence.

Why is evaluation evidence sometimes equivocal?

Cook and Spirrison (1992) reported on an evaluation of the Mississippi Project Aware program. One way to describe Project Aware was that it seemed to be a kinder and gentler type of “Scared Straight” program. Staff took juvenile delinquents to a maximum-security prison where they were exposed to an interactive ‘rap session’ with adult inmates. The inmates used the rap session to educate the juveniles about the travails of prison life and to share with them their own early experiences as juvenile delinquents. These rap sessions were delivered without the obscenity and verbal threats that characterized more aggressive versions of “Scared Straight” (e.g. Finckenauer 1982). But the explicit warning to the juveniles was the same: “clean up your act or you’ll end up here with us.” The Mississippi Project Aware Program sounds like a practical and common sense solution to juvenile crime. To Mississippi’s credit, they supported an evaluation to determine the program’s impact on subsequent delinquency and other measures.

In one part of their evaluation, Cook and Spirrison (1992) present before and after data for program participants. The group of kids who attended the rap session offered by Project Aware was measured at baseline on their mean number of new offenses as measured by official police contacts - before the program started - and later at 12 and 24 months. These means or averages were computed for the entire group of program participants. Cook and Spirrison note that the drop in criminality observed for the program participants was quite large: the average number of police contacts for the kids who attended the program dropped about 50% at 12 months. On the basis of these before and after data, it could easily be concluded that the Mississippi Project Aware program was a great success and should be widely implemented across the state. Reviews that included a number of evaluative studies like this could also come to the conclusion that “programs like Project Aware work.”
But the internal validity of the evaluation design (referred to as a simple pre-post or before and after test) is very weak because it does not rule out many of the potential rival explanations for improvement in the treatment group. For one, program participants may have improved over the 12- and 24-month period because they matured or aged. Another explanation for the improvement in the Project Aware kids is a type of selection bias called regression to the mean. Sometimes, kids are selected for programs like Project Aware because they commit offenses at very high rates. Even without intervention, most of these kids will return or regress to their nearly average rates of delinquent conduct. As Maltz and his colleagues (1980) note, regression to the mean can make ineffective programs look effective unless the evaluation controls for this effect. Thus, studies using such pre-post designs often result in very equivocal findings.

**Internal validity and quasi-experimental design**

One way to counter these threats to internal validity is to make the same before and after measurements on a comparable group of children not exposed to the treatment (Campbell and Stanley 1963). Evaluators using such an approach usually attempt to obtain a group that did not go through the program but is as similar as possible to the participants. In the Mississippi Project Aware study, the evaluators could have tried to obtain a comparison group of juvenile delinquents from another part of the state who were not exposed to the program. Non-equivalent comparison group designs such as this are often referred to as quasi-experiments (Cook and Campbell 1979).

The internal validity of quasi-experiments is much improved over pre-post designs because they can control for some internal validity threats such as maturation. It would control for maturation if we are confident that the comparison group of children not exposed to treatment was, in the aggregate, the same age and maturity level as their experimental group counterparts. Designs like this are called non-equivalent comparison group designs because it is assumed at the outset that the kids in the program will be different on some characteristics than the kids who participated in Project Aware. In such designs, no attempt is made by evaluators to control those pre-existing differences. This leaves open another possible explanation for the results we observe in quasi-experiments: selection bias.

Selection bias is a common problem confronted by evaluators in criminal justice. In many reports of quasi-experimental studies, the authors conclude that - whether the results are positive or negative for treatment - the observed outcomes may have been due to uncontrolled differences rather than any intervention effect. For example, in this hypothetical example in which the evaluator finds boys from another region in Mississippi to evaluate Project Aware, one could argue that preexisting differences between the experimental and comparison groups influenced the findings. For example, the comparison group could have come from the courts of a very impoverished part of Mississippi, in which the rate of delinquency is much higher. We would expect a comparison group of kids drawn from that area to be much more advanced and persistent in their delinquency (not susceptible to the “regression to the mean” problem) and therefore more difficult to
change with intervention. If this were true, then the evaluation would be biased toward showing that Project Aware worked. In other words, the kids in the comparison group could have committed offenses at a higher rate than the program participants. Because we did not control for this potential bias, any claim that the program “worked” because its participants had lower re-offense rates would be challenged.

We can think of selection bias as being comprised of two kinds: intentional and unintentional. Intentional selection bias is common for justice programs and other social interventions. Treatment and program providers sometimes select the kids they think have the best chances for success if exposed to their intervention. For example, they may select the oldest and more mature kids, or the ones with stable families, or those who were recommended by significant others such as teachers or police officers. These selections may lead to a ‘creaming’ effect, in which the kids with the best chance for improvement regardless of the intervention participate in the program. This observed improvement they show may then be mistakenly credited to the program rather than to bias in the selection of the kids for the comparison study.

Evaluators have developed a number of other methods for addressing selection bias and internal validity problems. The classic texts by Campbell and Stanley (1963) and Cook and Campbell (1979) outline some possible solutions. One method is known as matching. Evaluators using matched designs attempt to reduce selection bias by insuring that individuals in the comparison group are as similar as possible to the participants in treatment by matching individuals on particular factors they believe affect the outcome.

For example, researchers in the Project Aware study, if they were to conduct a matched design, would attempt to select juveniles in the comparison group who were as similar as possible on traits like age, race and prior record as the participants because these factors are considered predictors or important variables in whether a juvenile goes on to commit a new offense. Matched comparison groups provide stronger internal validity than many other types of comparison group designs. Although the potential for intentional selection bias is reduced, it does not provide a perfect solution. For example, investigators are generally able to match only on a few selected variables (e.g. Farrington, 1983).

Matching designs also assume that the researchers know in advance what variables are most important for matching and that data on those variables are available before the project starts (Taylor, 1994). To further complicate things, matching on some variables can cause mismatching on others (Sechrest and Rosenblatt, 1987). Going back to the Project Aware example, the researchers may have been able to match on age, race and prior record. Many other factors, however, have been theorized as predictors of future criminality, including criminal history of parents, birth order, school performance, to name just a few. In some situations, no data to examine the influence of these and other variables on the outcome are available. If the treatment group had more participants with stronger family support than comparison group kids, it could be that factor and not program participation that resulted in success for Project Aware kids.

These same issues also hamper our confidence in statistically controlled evaluations. Generally,
in contrast to matching techniques, evaluators apply statistical controls on a post-hoc basis, after the evaluation is over. In statistically controlled studies, the evaluator attempts to remove the influence of factors other than treatment participation that might explain the findings, with the help of quantitative methods such as multivariate regression. As with matching, the satisfaction that reviewers have when examining the statistically controlled study depends on a number of conditions. Two seem to be most critical: (1) that the data are available for all relevant and possibly explanatory factors to include in the statistical analysis; and (2) that the statistical procedure employed was appropriate given the factors controlled and the size of the sample in the dataset (i.e. number of participants or other units of analysis included in the evaluation).

Unintentional bias, on the other hand, refers to the possibility that the way that participants end up in a program is due to some yet unidentified variables. Because of these variables, program participants may make more improvement than the comparison group and not because of their exposure to treatment. For example, the evaluators of Mississippi’s Project Aware might have selected a comparison group and been able to match them to participants on a number of important factors, such as race, age, and prior record. But what if kids in the comparison group were still different after matching than the Project Aware participants in one very important but yet unknown way?

Whether intentional or unintentional, selection bias always offers the possibility that a result is being stacked in favor of one group or another. Questions of fairness, ethics, and importance stress the need to develop a fairer way of making the call whether the program worked or not.

Randomized field trials

Given that so many reviewers and commentators were critical of evaluation evidence and emphasized low internal validity as a major obstacle in determining what works, it is no surprise that many would recommend that randomized field trials be conducted whenever possible. Also known as randomized, classic, or true experiments, arguments for the use of field trials have been made for decades for testing various justice interventions, including offender treatment programs (e.g. Sechrest et al. 1979) and legal innovations (e.g. Federal Judicial Center 1981). Over ten years ago, the MacArthur Foundation Justice Study Group strongly urged the increased use of randomized trials and longitudinal studies to evaluate criminal justice programs and policies (e.g. Farrington, Ohlin and Wilson 1986). The University of Maryland Report to Congress also aggressively recommended randomized trials, including an ambitious plan to assess the effects of federal support for crime prevention by randomly allocating census tracts to different levels of funding (Sherman et al., 1997).

If a randomized trial is carried out with full integrity, we can assume with greater confidence that the program caused the observed changes in the outcome variable. Random assignment permits a stronger and less equivocal causal connection between the program and the observed outcome than other evaluative designs (e.g., Farrington, 1983).
This is because randomization is the only method that can theoretically control both known and unknown selection biases that often confound the interpretation of results. When evaluators employ randomization, they ensure that individuals are assigned to the program by chance probability alone. For example, the evaluator may use a random numbers table or a computerized randomization program to make the assignments. When using randomization, the evaluator insures that each individual has the same probability of being assigned to either a treatment or control group, what Garner (1977) calls equal probability assignment. By assigning individuals at random to experimental and control groups, extraneous and unknown variables that often cloud the interpretation of evaluations of criminal justice programs theoretically are balanced between the groups (e.g. Farrington, 1983). In other words, neither the treatment group nor the comparison group should have an advantage over the other on the basis of known or unknown variables. The groups should only differ on only one characteristic: participation in the program.

The Mississippi Project Aware study provides an illustrative example for us to consider. Cook and Spirrison (1992) did not just implement a pre-post evaluation or a quasi-experimental solution but conducted a randomized trial to test the impact of the program on subsequent re-offending and other outcome measures. The researchers wanted to insure that the kids who attended Project Aware - before the program began were as similar as possible to a group of kids who did not attend. The optimal way to insure was through randomization. Despite the term, there is nothing haphazard about the methods of random assignment: it is a systematic way of assigning group membership (Weisburd and Petrosino forthcoming). After determining that only juvenile delinquents who had been to the county court would be eligible for the study, the researchers randomly assigned 232 children to one of two groups: (1) to attend Project Aware; and (2) to a no-treatment control group. Whether they ended up in one group or the other was up to chance probability and not the wishes of the researchers or the desires of program staff. Unless some problem with the experimental study is revealed, we have greater confidence that any observed differences between the groups were due to the effects of Project Aware.

Recall that the before and after data revealed an approximate 50% decrease in the mean offending rates for the program participants. Because this was a randomized trial, the same before and after contrast was reported for the control group. The individuals assigned to this condition did not receive any special intervention or contact. The drop in mean offending rates for the control group was also approximately 50%! A pre-post study - without a control group - would have pointed to program success. Instead, the randomized trial demonstrated that the drop in criminality could not be attributed to the program but was likely due to some other factor that affected both groups simultaneously such as maturation or regression to the mean. The researchers therefore reported no difference between the kids in the Project Aware group and the control group in subsequent reoffending.

Randomized trials can potentially counter all of the internal validity threats raised by Campbell and Stanley (1966) and Cook and Campbell (1979) in their classic evaluation treatises. By including an equivalent control group not exposed to treatment, the design rules out maturation
or aging, as both groups will be aging or maturing. If a natural regression to the mean occurs, it should be observed for both groups, since both groups should have been demonstrably equivalent on baseline offending rates. Intentional selection bias is removed by the design, provided that treatment providers are not allowed to overrule or subvert the randomization process (e.g., Dennis, 1988). If implemented with full integrity, unintentional selection bias - the most difficult threat to control for in other designs - can also be ruled out (e.g., Lempert and Visher, 1987).

**Recognition of randomized field trials**

Randomized trials receive broad support among researchers and evaluators as the gold standard or Cadillac of research design because of their strength of internal validity. For example, Weiss refers to the randomized trial as the “classic design” (1972:9). It is likely due to this reputation that randomized trials are specifically recognized in solicitations for research funding proposals. For example, the 2001 U.S. National Institute of Justice solicitation for research in corrections and sentencing urged that randomized experiments be used whenever possible to evaluate the effects of an intervention.

The advantages provided by randomized trials are also implicitly recognized in rating systems designed by review teams to appraise evidence from primary or original studies. For example, in *The Effectiveness of Correctional Treatment* (Lipton, Martinson, and Wilks, 1975), a well-designed randomized trial received the highest score (“1A”). The University of Maryland *Report to Congress* graded a similar trial as “5” on a five-point scale of methodological rigor (Sherman, et al., 1997). Such rating systems explicitly interpret the evidence from randomized studies as being more trustworthy than those generated from other designs.

Several prior attempts to exclusively identify and collect controlled studies in criminal justice have been reported (e.g., Weisburd, Sherman and Petrosino, 1990; Dennis, 1988; Lemert and Visher, 1987; Farrington, 1983). Their activities also underscore the broad consensus that randomized trials are important. For example, Farrington (1983) located 42 field trials that tested interventions designed to ‘help people in the natural environment,’ or were conducted with the police, courts or in correctional institutions. He imposed a strict set of eligibility criteria: 1) the report had to have a clear statement of randomization; 2) use individuals and not aggregates as the unit of analysis; 3) a minimum of 50 participants assigned to each group; and 4) the results must have been formally published in a journal or book.

Dennis (1988) investigated implementation factors in successful criminal and civil justice experiments through written reports and scheduled telephone interviews with principal investigators. He located 41 experiments in criminal and civil justice conducted in the United States since 1972 that had at least 13 participants each in treatment and control groups. Dennis’ interviews were designed to compensate for missing information in original reports on issues like program and design fidelity. He found that the control of the randomization process by researchers rather than practitioners resulted in significantly lower rates of “covert manipulation,” the discretionary
overriding of the random assignment procedure without proper guidelines or researcher knowledge. Such violations can weaken the internal validity of an experiment.

Weisburd, Sherman and Petrosino (1990) compiled a *Registry of Experiments in Criminal Sanctions, 1950-1983*. The project began with the goal of locating, acquiring and reanalyzing the original data sets from experiments testing criminal sanctions already reported in the literature. Unfortunately, few investigators could provide their original data and the project did not advance beyond the acquisition of data sets already available in accessible archives. Given these and other obstacles, the investigators turned their attention toward producing a registry of 76 experiments with an electronic data set of study characteristics as reported in the original evaluation documents (Weisburd, et al. 1990).

**Criminological experiments in the U.S.: a brief history**

The first criminological trial was likely the Cambridge-Somerville Youth Study (Powers and Witmer, 1951). In this experiment, investigators first matched individual participants (i.e. youths nominated by teachers or police as 'troubled kids') on certain characteristics and then randomly assigned one to a counseling group and the other to a no-treatment control. Investigators have consistently reported that the counseling program had no positive impact and likely caused more harm than doing nothing (McCord, 1978). Although the first major report did not get published until 1951, the first child was likely assigned randomly sometime during 1937.

Randomized trials were used selectively in criminology and elsewhere until the 1960s. Interest in such studies was stimulated after the publication of *Experimental and Quasi-Experimental Designs for Social Research* (Campbell and Stanley, 1963). Evaluators were now armed with an arsenal of tools for implementing randomized and quasi-experimental designs, but the book clearly recognized a hierarchy of techniques based on how well they controlled internal validity (Campbell and Stanley, 1963). Because of their high internal validity, randomized trials soon became an important research design across the social sciences in the 1960s (Oakley, 1998). This was particularly true in the United States, as they were frequently used to evaluate social programs during President Lyndon Johnson's effort to launch a 'great society.' Oakley (1998) notes that randomized studies fell out of favor and were less frequently used after the 1960s because they consistently questioned the efficacy of government programs.

Randomized studies again appeared on the radar screen of criminologists in the mid-1980s. Some credit for that should go to David Farrington, Lloyd Ohlin and James Q. Wilson (1986) for their influential book, *Understanding and Controlling Crime*, which recommended the use of randomized studies whenever possible to test criminological interventions. Another influential factor was the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984). In this study, the investigators were able to get the police to cooperate with a design that took away police discretion and instead randomly assigned misdemeanor domestic violence cases at the scene to one of three conditions: arrest, mediation, or separate suspect and victim for 8 hours.
The investigators reported that arrest reduced subsequent violence in a six-month follow-up (Sherman and Berk, 1984). This study not only influenced the adoption of mandatory arrest policies by police departments across the nation but also increased the visibility and legitimacy of randomized trials in criminology.

In the 1980s, the National Institute of Justice embarked on an ambitious plan to fund criminal justice trials, including the Minneapolis study and several replications (Garner and Visher, 1988). The Institute's funding included the sponsorship of a special working group and a workshop on field experimentation in criminal justice, under the auspices of the National Research Council's Committee on Research on Law Enforcement and the Administration of Justice (Lempert and Visher, 1987). Its goal was to promote the sharing of perspectives, disseminating results and insights from major trials, discussing issues that arise in conducting such studies, and providing a forum for the exchange of views. Whether these goals were successfully reached at the workshop is not known, but Garner and Visher (1988) note that 24 experiments were funded by NIJ during the next year alone.

Finally, the Academy of Experimental Criminology was created in 1998 to promote the use of randomized trials in justice settings (http://www.crim.upenn.edu/aec). One way it will promote studies is by honoring as Fellows those who conduct randomized field trials or promote their use in other important ways.

**How do randomized trials compare with non-randomized studies?**

Despite the recognition and broad support for randomized trials, comparative research that has examined how well randomized field trials compare with other designs can come to diverging results depending on the type of empirical study. Generally, there are two main types of comparative studies of field trials with other designs. The first compares estimates from an experimental design with non-experimental methods in the same research setting, evaluating the same program. In this case, randomized trials seem to come to very different conclusions than non-randomized designs, and a strong case for a design effect can be made.

For example, Lalonde (1986) compared the results of econometric analyses with experimental estimates of program impact for an employment program. This was an important comparison because econometricians sometimes argue that they can approximate a random assignment study by statistically controlling for the influence of extraneous factors except the intervention. Lalonde (1986) instead reported that the econometric analyses grossly overestimated the benefits of the employment program when compared to the more modest results provided by the experimental design.

Another type of statistical matching procedure known as propensity scores has been offered as an optimal way to approximate randomized designs in quasi-experimental studies. In a comparative study of evaluations in school-drop out programs, researchers found that the estimates of program benefit from the propensity score study versus the randomized design differed dramati-
cally across four sites (e.g. Agodini and Dynowski, 2001). Some believe this is the most appropriate comparison for such "design studies" as they rule out some of the influence of the program, setting, timing, and other variables on the results.

The other major type of comparison is made by examining the results for randomized studies versus other designs in large syntheses and reviews. The motivation behind these studies is that if the type of design has an effect, it should be observed when randomized studies are compared to non-randomized studies across scores, even hundreds, of studies. In such studies, the case for a design effect is less clear.

For example, Lipsey and Wilson (1993) reported an analysis of over 300 quantitative reviews relevant to psychological, educational and behavioral treatment. They reported that randomized and quasi-experimental evaluation across a large number of meta-analyses produced very similar estimates of program impact. Only simple pre-post studies (without any control or comparison group) consistently reported results much larger than randomized or quasi-experimental designs. Shadish and his colleagues (1996) also examined meta-analytic data and reported that randomized and quasi-experimental studies seem to come to fairly similar conclusions across an aggregate group of studies. This was especially true of quasi-experiments that demonstrated group equivalence on pretest or baseline measures.

On the other hand, Kunz and Oxman (1998) examined estimates of randomized trials and non-randomized studies in health care and found that the estimates from non-random studies varied widely from those reported in randomized ones. Weisburd and his colleagues (2001) used the studies annotated in the University of Maryland Report to Congress (Sherman, et al., 1997) to test for a design effect. They found that randomized studies, even when compared to well-implemented quasi-experimental studies, were less likely to report program success and more likely to find the program was harmful to its recipients (c.f. Bailey, 1966). Although a rigorous review of these 'design studies' is needed, these finding should inspire caution before assuming that non-randomized studies approximate results from randomized field trials.

**When problems compromise the internal validity of experiments**

Randomized experiments are criticized on a number of grounds, including the ethics of randomly assigning persons committing similar offenses to different punishments through random assignment. There are also concerns about whether the findings from an experiment can be generalized to other settings, a quality known as external validity. These are just a few of the many criticisms of experiments and researchers must consider them before implementing a randomized trial. But for reviewers attempting to determine what works with regard to justice programs by collectively summarizing a number of studies, ethical or external validity problems of experiments are not focal concerns. The relevant issue is whether the experiment experienced any kind of problem that threatens the confidence that can be placed in the observed findings. The three major issues that reviewers generally confront when examining original reports of experiments
are: (1) whether the random assignment protocol was violated in any way; (2) whether there was substantial attrition or loss of subjects from the study; and (3) whether the treatment program was implemented to a degree sufficient to distinguish its outcomes from that of the control condition.

Calling a study an "experiment" does not guarantee that the study has strong internal validity. Things can go wrong to undermine the experiment. Although reviewers are generally concerned with what goes wrong, such issues are usually not reported in sufficient detail to take into account in the review. Deficiencies in report writing are what led Dennis (1988) to attempt to interview the original experimenters to fill in the missing gaps in the study reports. Nonetheless, reviewers look for clues in the original documents that the experiment experienced no compromise of randomization, had little or no attrition, and kept treatment and control conditions distinct. Random assignment violations. Randomization is the key component in the experimental design and researchers consistently recommend that steps be taken to insure the integrity of the assignment protocol (e.g. Boruch 1997). Even when treatment practitioners, for example, have agreed to the conditions for the experiment, they still may have a vested interest in overriding or violating the protocol in order to assign certain participants to certain groups (e.g. Dennis, 1988).

For example, in a Denver experiment, Ross and Blumenthal (1974) were able to persuade sentencing judges to randomly assign drunk-driving offenders to one of three conditions: probation, counseling or a monetary fine. The experiment was agreed to by the judges because they often implemented all three dispositions for drunk driving offenses anyway but had little rationale for choosing one sentence over another. Despite the consent of judges, the experiment had problems when implemented in the field. Ross and Blumenthal (1974) detail how judges that subsequently violated the assignment procedures on a frequent basis, usually at the bequest of the defense attorney who objected to a client being randomly assigned to conditions such as counseling or probation, both of which were considered by clients to be more onerous than simply paying a fine and leaving. Thus, even with the best of intentions, a randomized field experiment can go awry because of objections of legal practitioners to assignment of their clientele to a presumably more punitive condition.

Attrition. Attrition also threatens the conclusions from an experimental design. Attrition is the loss of subjects from study groups after randomization. This is a problem in any longitudinal study - experiments included - that follow participants for some period of time after exposure to treatment (e.g. Farrington 1983). Attrition can affect confidence in the results obtained because it can change the composition of the groups. For example, if there is a loss of participants from the study after one year in the community (e.g. through death of the participant, inability to locate the individual’s records), the groups - assumed to be equivalent immediately after randomization - may no longer be so. This threatens internal validity if the participants who drop out from the treatment group differ in some important way from those who dropout of the control group.

The distinctiveness of treatment and control conditions. The purpose of doing an experiment...
is to test the intervention’s effect on crime relative to another condition. For this to have meaning, the treatment must differ from the condition it is compared to. Otherwise, the results have no meaning. Although it sounds obvious, it is sometimes difficult to maintain the distinctiveness between experimental and control groups. This is especially true in experiments in which the only difference between the groups is the “amount of contact” that participants have with staff. For example, an experiment may randomly assign probationers to an experimental condition in which he is to receive two phone calls per week by the probation officer supervising him. The control condition may involve only a single phone contact. In experiments like this one, it is possible for the experimental condition to be less intense and the control condition to become more intense during the experiment so the end result is that both groups received an average 1.5 contacts per week. The experimental comparison would be meaningless if the groups were indistinguishable on the treatment variable.

Conclusion

In this paper, I have discussed the role of experiments from the perspective of the reviewer. Over a four-decade history, reviewers attempting to find out what works by summarizing available evaluations have continuously criticized the quality of the studies, particularly their internal validity. Consequently, many of these reviewers have urged that more experiments be conducted to test for the relative effects of justice programs whenever possible. I summarized the strengths that randomized experiments present and concluded the paper with a discussion of some of the major problems that could threaten the internal validity of an experiment. These problems should not be overstated and can sometimes be minimized with careful planning and cooperation of policymakers and practitioners in the field setting.

Acknowledgements

This paper was supported in part by a Mellon Foundation grant to the Center for Evaluation, American Academy of Arts & Sciences, a Smith-Richardson Foundation grant to the Jerry Lee Center of Criminology, and a National Institute of Justice grant to the Rutgers, School of Criminal Justice. Professor Frederick Mosteller was most generous with his time, encouragement, and editorial criticism. The paper is certainly better because of Fred. It was also improved by comments from anonymous peer reviewers, and from earlier conversations with - and comments from - David Weisburd, Todd Clear, Jim Finckenauer, Ron Clarke, and Paul Lerman.

Notes
1 I use Cook and Spivison’s (1992) study as an example to illustrate the strength of randomization. We note that their experiment encountered other problems including attrition. The losses included 19 from the experimental group and 37 from the control group.
2 “Scared Straight” is the more popular name associated with the Rahway State Prison Lifer’s Program in New Jersey. In the Lifer’s program, inmates attempted to scare delinquents or high-risk youths from a life of crime by a brutal rap session in which they described the horrors of prison life (Finckenauer, 1982).
The mean number of offenses is the total number of offenses committed by each person over the study period, divided by the number of persons in the group. If 10 persons each commit one offense over a six-month period, the mean offense rate is 1.0 over the six months (10 x $\frac{1}{10} = 1.0$).

References


Cook, Thomas D. and Donald T. Campbell, 1979, Quasi-experimentation: Design and Analysis Issues for Field Settings, Rand McNally.


Lipton, Douglas, Robert Martinson, and Judith Wilks, 1975, The Effectiveness of Correctional Treatment, Praeger.


JAPANESE JOURNAL OF SOCIOLOGICAL CRIMINOLOGY No.30 2005


Online source, http://www.upenn.edu/aeec, last checked on April 8, 2005.


Email: anthony_petrosino@hotmail.com

70  課題研究 最近の刑事政策関連立法・施策における政策形成過程の再検討
司法プログラムを評価するための米国における
無作為化フィールド実験の利用
——概論——

アンソニー・ベトロシーノ
リサーチコンサルタント


キーワード：実験デザイン、無作為割付け、プログラム評価

E-mail: anthony_petrosino@hotmail.com