1520 H Street, N.W.
Washington, D.C. 20005
Telephone 202/633-6011

EXPERIMENTATION IN THE LAW

Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law

Federal Judicial Center
September 1981

This publication is the report of the Federal Judicial Center Advisory Committee on Experimentation in the Law, which was established by the Federal Judicial Center and charged to make such study and recommendations as the Committee deemed appropriate. This report was issued by the Committee on March 22, 1981. The analysis, conclusions, and points of view are those of the Committee. It should be noted that on matters of policy the Center speaks only through its Board.

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402

Cite as Experimentation in the Law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law (Federal Judicial Center 1981).

TABLE OF CONTENTS

PREFACE

In January, 1978, Chief Justice Warren E. Burger, as Chairman of the Board of the Federal Judicial Center, appointed the Federal Judicial Center Advisory Committee on Experimentation in the Law "to identify, define, analyze, and recommend resolution of issues bearing on the propriety, value and effectiveness of controlled experimentation for evaluating innovations in the justice system" and "to provide guidance to researchers, judges and administrators who must decide what areas are appropriate for controlled experimentation." These terms of reference are contained in a letter from the Chief Justice to the committee's chairman, Chief Judge Edward D. Re, dated January 24, 1978. This letter is reprinted as Appendix A.

The committee was established because of questions about whether the operation of the justice system could be improved through empirical research without violating important ethical values. Experimental programs have been undertaken at an increasing pace, but evaluation often has not produced clear conclusions about the achievements of these initiatives. Many observers have recognized the "true controlled experiment" or "randomized experiment" as a powerful tool for evaluating innovative programs and procedures, but its use in the justice system raises ethical and legal questions.

The Judicial Center recognized a need to improve the evaluation of experiments conducted within the justice system, and that scientifically rigorous evaluation methods might meet that need if and when they were acceptable. It also recognized that responsible arguments could be made both for and against methods such as that of the randomized experiment.

The Chief Justice therefore undertook to appoint a committee to address and suggest resolution of these issues. Two goals guided the appointments. First, the committee membership should include the rather broad spectrum of interests and disciplines pertinent to the task. Relevant fields of study include constitutional law, research methodology, legal ethics, and the ethics of research involving human subjects. Specially relevant interests include those of the bench, the bar, program administrators, and potential experimental subjects. Second, the committee should be as free as possible

from preconceptions that would incline it to favor or oppose experimentation in the justice system. Although the Center's Research Division provided financial and staff support for the committee, it had no say in the committee's decisions. No member of the committee had any prior special commitment to research or experimentation in the justice system.

At its first meeting, the committee resolved to ask the Department of Justice and the National Center for State Courts to designate advisors to serve as nonvoting members, in order to broaden the committee's base of knowledge about actual research activity in the justice system. In addition, three of the committee's early meetings were devoted to consultation with experts in social science research about the scientific debate and consensus regarding evaluation methodology.

The committee's deliberations continued through ten two-day meetings. The subject matter has presented and continues to present a rich challenge that this report only begins to address. It is the committee's hope that it has provided a fruitful starting point for illuminating the need for and proper role of experimentation to aid policy decisions in the administration of justice, as well as in related areas of public administration.

It would be difficult to exaggerate the contributions made by the committee's staff in the preparation of this report. The committee, representing diverse disciplines and orientations, has had to consider difficult and complex matters. In attempting to move from basic concepts to specific guidelines for justice system experiments, we relied heavily on the staff to gather and present for our consideration the issues and perspectives of experts in the relevant fields.

John Shapard deserves special recognition for his sustained contributions. He served throughout the project as our secretary, principal staff person, and general assistant. He prepared a number of reports for our early meetings that surveyed the questions with which we were to deal, made numerous contributions to our analysis in the course of our meetings, distilled the consensus from our evolving debate, and was of great assistance in connection with drafting of the report at all stages. The wise and experienced counsel of William B. Eldridge was also of invaluable assistance at critical points in our deliberations.

The Committee:

Honorable Edward D. Re, Chairman
Chief Judge, United States Court of International Trade

Alvin J. Bronstein, Esquire
Executive Director, The National Prison Project of the American Civil Liberties Union Foundation

Alexander Morgan Capron
Professor of Law, University of Pennsylvania, on leave as Executive Director, President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research

Honorable Wilfred Feinberg
Chief Judge, United States Court of Appeals for the Second Circuit

Jane Frank-Harman, Esquire
Manatt, Phelps, Rothenberg and Tunney
Washington, D.C.

Paul A. Freund
Carl M. Loeb University Professor Emeritus, Harvard University

Gerald Gunther
William Nelson Cromwell Professor of Law, Stanford University

Alasdair MacIntyre
University Professor of Philosophy and Political Science, Boston University

Norman Redlich
Professor of Law and Dean, New York University School of Law

Jerome J. Shestack, Esquire
Schnader, Harrison, Segal and Lewis
Philadelphia, Pennsylvania

Honorable Joseph T. Sneed
United States Court of Appeals for the Ninth Circuit

Honorable Abraham D. Sofaer
United States District Court for the Southern District of New York

June Louin Tapp
Professor of Child Psychology and Criminal Justice Studies, and Adjunct Professor of Law, University of Minnesota

Advisors:

Joel Zimmerman served as advisor from the National Center for State Courts in his capacity as Research Director of the National Center's Programs Division

Preface

Honorable Daniel J. Meador and
Honorable Maurice Rosenberg served as advisors from the United States Department of Justice during their respective terms as Assistant Attorney General, Office for Improvements in the Administration of Justice

Staff:

(All are members of the staff of the Federal Judicial Center)

William B. Eldridge, Director of Research

Gordon Bermant, Deputy Director of Research

E. Allan Lind, Research Associate

John E. Shapard, Research Associate

CHAPTER I. INTRODUCTION AND SUMMARY

A. Introduction

Experimentation has a long and important history in our system of justice and in public policy generally. A great strength of our society is that we are open to innovative methods for solving problems and are willing to accept the diverse approaches of various states, communities, and authorities within our federal system of government. Justice Brandeis, dissenting in *New State Ice Co. v. Liebmann,* expressed this spirit eloquently:

> [A]dvances in the exact sciences and the achievements in invention . . . [i]n large measure . . . have been due to experimentation. . . . It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country. This Court has the power to prevent an experiment. . . . But in the exercise of this high power, we must be ever on our guard, lest we erect our prejudices into legal principles. If we would guide by the light of reason, we must let our minds be bold.[1]

Experimentation is an effective tool for improving the administration of justice. Achieving the goals of the justice system—which include preventing as well as punishing crime, ensuring justice in civil and criminal cases as expeditiously as possible, and reducing the costs of providing and obtaining justice—requires that the system be flexible and willing to adopt new programs and procedures. Although our Constitution and laws establish certain basic procedural guarantees that are not readily subject to modification, other features of our justice system are more open to change. Among these are programs and institutions administered by prisons, courts, and probation agencies, for example. These are elements of the administrative structure of justice—the particular means for affording procedures guaranteed by the Constitution, such as trial by jury and due process of law.

1. 285 U.S. 262, 310-311 (1932).

In considering and evaluating the various methods of achieving justice, one must be concerned with the effectiveness of existing procedures, programs, and institutions compared with the potential effectiveness of available alternatives. Proposed innovations are frequently of uncertain value, for it is often unclear whether they will result in the improvements they are intended to achieve, or will do so at acceptable costs and without unacceptable adverse consequences. Sometimes, reasoned judgment based on available information and experience will be sufficient to determine whether a proposed innovation should be adopted or not. Often, however, uncertainties regarding either the risks of adverse consequences or the possibility that the innovation will be ineffective make it impossible to reach a rational judgment without additional information.

When available information is inadequate, how are these uncertainties to be resolved? The answer will often be: only by some form of experiment that permits a comparison between the results of the proposed innovation and those achieved by the existing method of pursuing a given goal. The controlled, *i.e.*, randomized, experiment is the form that permits the most reliable comparison. However, because the nature, value, and ethical acceptability of controlled experimentation cannot be assessed without reference to other research methods, we include in this report a discussion of issues involved in justice system experimentation by methods other than that of the randomized experiment. Without experimentation in some form, it will often be impossible to evaluate an innovation adequately before it is implemented.[2]

The need to evaluate proposed innovations through scientific experiments has been recognized increasingly in recent years, and a number of highly informative experiments have been conducted within the justice system. But the relationship between science and the administration of justice has been tenuous. All too often, either innovation has proceeded without needed prior experiment, or experiments have been undertaken without enough forethought about whether and how they will provide the required information. In other instances, experiments have been undertaken with zeal for both precision and clarity of results, but without corresponding attention to the relevance of the results. Thus, chances have been missed to evaluate reliably the effectiveness of alternative methods for administering justice.

2. Except where the context suggests otherwise, the term "experiment" is used throughout this report in its general, popular meaning to refer to a test of a new concept or program. This report does not generally use the term "experiment" in its narrower technical sense, which is synonymous with "randomized experimental design," as defined at page 17, *infra*.

Moreover, poorly designed experiments may be worse than no experiments at all. They may lead to unjustified faith in the merits of innovations and to unjustified lack of faith in the value of experimentation. They may expose individuals involved with the justice system to unwarranted risks or harms and deprive them of the respect and principled treatment that is their due.

Finally, even well-designed justice system experiments may raise ethical issues. Indeed, the more rigorously designed the experiment, the more risk that it could create significant disparities among subjects or could deceive the subjects about its true nature or intent.

Recognizing the complexity of these problems, Chief Justice Warren E. Burger appointed the Federal Judicial Center Advisory Committee on Experimentation in the Law. The committee's mission was to recommend ways to ensure that experimentation within the justice system proceeds soundly, in a manner that will advance the cause of justice both in the means employed and the ends achieved.

This report addresses what the committee has called "program experiments" within the justice system. A program experiment is an alteration in the actual operation of the justice system designed to show whether such an alteration would be an improvement over the status quo. Program experiments are sometimes used to determine whether an existing program should be abolished, or whether one or more existing programs should be employed in a new manner. Usually, however, program experiments involve limited implementation of an innovative program. Any practice, rule, procedure, law, or policy carried out as part of the administration of justice can be considered a "program" for our purposes. Other kinds of research and experimentation can sometimes inform decisions about the effectiveness of proposed innovations in the administration of justice, and can therefore have an important role in improving the operation of the justice system. This committee's mandate, however, is limited by the distinctive feature of a program experiment—experimental change in the actual operation of the justice system—which has as a necessary consequence some direct influence on the interests of individuals involved with the justice system.[3] Hence, although a program experiment might produce

3. Program experiments are distinguished from research that does not directly alter the operation of the justice system, including, in particular, experiments that merely simulate or test by analogy. Simulation experiments, such as one in which new jury instructions are tested by obtaining verdicts from jurors who view a simulated trial, may raise ethical questions of their own. These questions are best considered as part of a general discussion of the ethics of experimentation upon human subjects. The report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research discusses them usefully and construc-

basic insights into human capabilities, attitudes, or behavior which are valued for their own sake, its purpose is not to be found, or its justification sought, on that basis.

Throughout this report, we emphasize that a program experiment must be evaluated not in isolation, but in a double perspective: as a segment in the ongoing process of administration of justice and the effort to improve it, and as one of the choices available to justice system administrators, who must compare the alternatives open to them.

When an innovation is proposed in the administration of justice, those considering the change have three choices. First, they may simply retain the existing practice and forgo the innovation. But often the innovation will have been proposed precisely because the present practice is thought to be seriously inadequate, perhaps even a source of injustice. Maintaining the status quo may thus be undesirable. Second, they may adopt the innovation on a general basis without prior testing. But there will often be serious uncertainties about whether the innovation, although promising in theory, will in fact produce the desired improvements without undesirable consequences. Adopting the innovation in the face of such uncertainties would thus also be questionable. Third, they may adopt the innovation on an experimental basis (*i.e.,* undertake a program experiment) to resolve the uncertainties and thereby permit a more informed future choice between the first two options. But a program experiment that is effective in resolving the uncertainties may require an experimental design that itself creates problems of legal and ethical dimensions.[4]

Experiments must be designed to avoid misleading results. But that requirement may lead to practices that raise serious ethical problems because of the ways in which individuals involved in the justice system are categorized and treated. Such practices include providing the innovative program or treatment to some persons while providing the present (status quo) treatment to others who have the same or similar relevant characteristics; acquiring information that is normally private; and concealing from participants certain information about the experiment, to ensure that the experiment accurately predicts what would happen if the innovation were adopted on a nonexperimental basis. Even when an innovation is applied only to those who consent, not all individuals with

tively (Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, 44 Fed. Reg. 23,192 (1979)).

4. A fourth option may be available: retaining the present practice while a simulation experiment or some other form of inquiry is undertaken to resolve uncertainties. This preliminary course can sometimes be pursued before a choice must be made about actually changing the operation of the justice system.

similar relevant characteristics might be allowed to participate in the experiment.

Experiments within the justice system will often unavoidably involve compulsory participation on the part of individuals because the justice system has many compulsory aspects. The clearest example is an innovation that is intended for mandatory application; an experiment involving voluntary participants might not predict the effects of the program when mandatorily imposed. If, for instance, the innovation under consideration is a change from voluntary to mandatory pretrial conferences, it might be impossible to design a useful test involving voluntary participants.

Because program experiments in the justice system often involve mandatory participation, they create possible conflicts with societal and legal commitment to certain fundamental ideals. These conflicts must be resolved in a manner consistent with our legal and ethical norms. Disparate treatment of individuals must be reconciled with the constitutional requirement that differences in the treatment of similarly situated persons be justified. Likewise, experimental methods that compromise privacy or the obligation of candor must be justified in accordance with the importance of these concepts in our system of justice under law.

This report does not seek to discourage experimentation, but rather to foster responsible experimentation within the justice system. Because program experiments often generate conflicts among fundamental principles of our system of justice, however, the decision to experiment demands the kind of careful analysis and precautionary procedures recommended in this report.

An effectively designed program experiment can have ethical justification stronger than its ethical shortcomings. The alternative to such an experiment may often be a choice between continuing a present practice that is seriously flawed, or adopting a proposed innovation generally and risking the possibility that it may be worse than the present practice. Without sound experimentation, it may never be discovered that we have rejected an innovation that would have advanced the ideals of justice, and it may be discovered too late that we have adopted an innovation that undermines those ideals.

The committee recognizes that a sound experiment will sometimes require the most rigorous scientific methods, which may involve disparate treatment, intrusions on privacy, and less than total candor. Less significant, but still deserving concern, are the sometimes substantial cost and inconvenience associated with conducting such an experiment. Accepting these practices and consequences temporarily in order to obtain necessary information will

sometimes be preferable to accepting the risks associated with alternative courses of action.

In approaching its task, the committee has focused primarily on ethical considerations. Legal and constitutional principles set outer limits on what may be permitted within the administration of justice. But not all experiments that might be legally or constitutionally tolerable will also be acceptable on ethical grounds. The committee's task was to address officers of the justice system regarding experiments they might undertake as administrators, in contrast to experiments whose validity they might have to decide in their capacities as judges, counsel, or other legal officers. Therefore, the committee's work focuses on whether an experiment is justifiable according to ethical principles perceived as fundamental to our system of justice. Although this approach, with its emphasis on balancing competing ethical claims, may produce more restrictive standards than would a purely legal and constitutional analysis, the relationship of these ethical principles to fundamental constitutional precepts will be apparent.[5]

The recommendations offered in this report are neither intended nor suited to be adopted as strict standards for justice system administrators. The questions addressed are novel and often complex. They cannot be resolved adequately at first impression or abstractly. They must be decided on a case-by-case basis, and they must ultimately be resolved by the administrator under whose authority a program experiment is to be undertaken. The committee does not offer rules that prescribe what may or may not be done in program experiments; rather, it suggests an approach to analyzing those questions that will help ensure responsible answers. What factors need to be considered? What factors should not be considered? How do permissible considerations relate to each other? And, finally, what kinds of arguments will support a decision to undertake or not undertake a particular program experiment? Determining whether and how particular standards apply to a particular experiment must remain part of the judgment committed to the administrator. A more comprehensive set of principles will arise from the accumulated record of case-by-case judgments, and in Chapter VII, the committee recommends procedures for developing an accessible body of such judgments.

5. The issues presented by program experiments have received some constitutional scrutiny, but have yet to be faced squarely and thoroughly by the courts. That process has just begun. *See, e.g.,* Aguayo v. Richardson, 473 F.2d 1090 (2d Cir. 1973); United States v. Thompson, 452 F.2d 1333, 1339 (D.C. Cir. 1971); Department of Motor Vehicles v. Hardin, 58 Cal. App. 3d 936, 130 Cal. Rptr. 311 (1976); People v. Colon, 29 Cal. App. 3d 397, 105 Cal. Rptr. 695 (1972).

B. Summary

This report recommends that in deciding whether or not to undertake a program experiment, the decision maker consider four questions:

1. Do the circumstances justify consideration of a program experiment?
2. What experimental designs will be adequate to produce the required information?
3. What ethical problems might these experimental designs present, and how can they be resolved?
4. What authority and procedures are necessary for undertaking the experiment?

The first three questions address the general issue of whether a proposed program experiment would be justified according to certain basic ethical principles. The fourth question concerns both the authority of the administrator to undertake the proposed experiment and the procedures that should be followed in order to ensure that all four questions are adequately addressed.

Chapter II addresses the first question. Because experimentation within the operation of the justice system presents unavoidable ethical difficulties, program experimentation should only be considered when certain threshold conditions are met. First, the status quo must in fact warrant substantial improvement or be of doubtful effectiveness. Second, there must be significant uncertainty about the value or effectiveness of the innovation. Third, information needed to clarify the uncertainty must be feasibly obtainable by program experimentation, but not readily obtainable by other means. And fourth, the information sought must pertain directly to the decision whether or not to adopt the proposed innovation on a general, nonexperimental basis.

If these threshold conditions are met, the second question is reached: "What experimental designs will be adequate to produce the required information?" Chapter III briefly presents the theory and methods of experimental research design, illustrating the ways in which different types of experiments may, or may not, yield sufficiently precise and unambiguous results. An understanding of experimental design helps highlight ethical problems that emerge from such designs. In addition, such an understanding reveals that experimental design is not merely the technical concern of researchers, but is a crucial ethical consideration in the decision to undertake a program experiment.

Chapters IV through VI address the third and most complex question: "What ethical problems might a particular experiment

present, and how can they be resolved?" Chapter IV sets forth basic ethical principles that the committee has employed as a framework for its analysis. Two principles, equal treatment and respect for persons, are recognized as having paramount importance in evaluating experiments in the justice system. To the extent that experimental practices encroach upon these principles, they harm the interests of individuals, and must therefore carry a commensurate burden of justification.

The necessary basis for justifying infliction of harm is the benefit likely to be achieved. In weighing harms against benefits, one must recognize the varying significance attached to different kinds of harm or benefit. The crucial decisions about program experiments involve weighing harms to particular individuals against benefits to some larger group or to the general public. An essential standard for evaluating program experiments requires that individuals may be exposed to some particular harm or risk only when (1) some particular benefit can be achieved in no other way, and (2) the benefit to be achieved clearly outweighs the harm or risk. Program experiments often include potential benefit for the experimental subjects they harm; they may also involve risk, harm, cost, or inconvenience to the general public. The weighing of harms and benefits must take account of these factors as well.

The balance of harms and benefits associated with a proposed program experiment must always be evaluated in light of the harms and benefits that may ensue from alternative courses of action—retaining the status quo or innovating without prior experiment. Furthermore, the balancing process is constrained by absolute limitations: some harms to individuals cannot be outweighed solely by benefits to others, no matter how great those benefits may be.

Chapter V applies the general ethical principles described in Chapter IV by evaluating the kinds of harms that program experiments may entail and suggesting the level of probable benefit necessary to justify those harms. The following principles emerge:

1. Mandatory imposition of harm poses the most difficult problems; risks or harms that responsible subjects freely accept will rarely require the degree of justification demanded when similar harms or risks are mandatorily imposed. Mandatory use of persons as means for experimentation is a separate harm, even when the experiment involves no harmful disparity.

2. Disparate treatment of persons involved with the justice system creates potential for harm and must be evaluated with particular concern for (a) the significance of the interests affected; (b) the extent of the difference between treatments; (c) a comparison of the disparity with standard treatments or expectations; (d) the degree to which the disparity reflects differences in qualification of subjects; (e) whether the experimental treatment is harmful or beneficial to the subjects; and (f) whether participation is mandatory or voluntary.

3. When an experiment risks infringing subjects' privacy, the risk not only carries a burden of justification, but also triggers an obligation to protect the confidentiality and, where possible, the anonymity of subjects.

4. Finally, even if the research process might be strengthened by concealing from the subjects that they are involved in an experiment in the justice system, or the nature of the experiment, concealment is a doubtful course and imposes a special burden of justification.

Chapter VI offers guidance on the central question whether or not ethical difficulties associated with a proposed experiment are justified—or can be justified—in light of benefits likely to be obtained. The chapter illustrates the kind of analysis that may properly be employed in certain recurring situations. For example:

1. It is possible to justify experiments involving very serious harms to individuals, but only when alternative courses of action—retaining the status quo or adopting the innovation without experiment—involve risking harm that is more significant than harm risked in the experiment.

2. Experiments involving less serious harm may be justified more easily, but only if they are the least harmful way to resolve satisfactorily the uncertainties that led to considering program experimentation.

3. Some experiments may be impossible to justify, because their harm to subjects exceeds outer limitations on what may ever be justified by benefits to others.

Chapter VII addresses the fourth question: "What authority and procedures are necessary for undertaking a program experiment?" The chapter analyzes the elements of authority that should exist as a precondition to undertaking program experiments, and then suggests procedures to ensure that such authority does, in fact, exist.

Two conditions are necessary to ensure adequate authority for undertaking a program experiment. First, the administrator under whose authority the experimental program is to be implemented must have legal authority to adopt the program on a nonexperi-

mental basis. Second, approval by an officer or body with a broader public mandate may be needed before proceeding with the experiment, because the administrator's authority to undertake the program may not necessarily encompass the kinds of harms the experiment entails. A particular program may be quite clearly within the authority of a probation official, for example, but mandatory imposition of that program on a disparate basis may exceed the official's mandate.

Chapter VII also recognizes that procedural mechanisms may be needed to ensure: (1) that experiments are approved by authorities with sufficient public mandate, and (2) that the ethical analysis recommended is undertaken and documented to aid future decisions regarding experiments. To these ends, the chapter recommends:

1. That advisory bodies be created within the various institutions of the justice system to offer guidance to administrators on matters of ethical analysis and experimental methodology, and provide appropriate approval of experiments involving ethical problems that may appear to exceed the sponsoring administrator's mandate, and
2. That the justification for an experiment be reported in writing by the responsible administrator, and that these reports be made available as informal precedents in the field of program experimentation.

The ethical problems of program experimentation deserve continuing attention and sensitivity. Decisions with regard to experiments must also consider the consequences of inadequate experimentation or of innovating without prior experimentation. Responsible experimentation, conducted with sensitivity to fundamental principles of justice, can be an important tool to improve the justice system. It is our hope that the analytical framework suggested in this report will assist administrators to use experiments to their fullest extent, consistent with the ethical standards that we propose as guides for decision.

CHAPTER II. CIRCUMSTANCES IN WHICH PROGRAM EXPERIMENTATION SHOULD BE CONSIDERED

The value of program experimentation should not obscure the practical and ethical difficulties that such experimentation almost inevitably entails. These difficulties will be considered in detail in later chapters. But their existence requires us to recommend initially that certain threshold conditions must normally be met before a decision to experiment is considered.

First, the present practice must either need substantial improvement or be of doubtful effectiveness. Even though an experiment may promise to yield valuable information about the proposed innovation, committing resources or risking the harms associated with experimentation will be difficult to justify unless there is a genuine need for improvement. Experiment for experiment's sake has no place in the justice system.

Second, there must be significant uncertainty about the value of the proposed innovation. Recall that when an innovation has been proposed as an alternative to some present practice in the administration of justice, three choices are open: to adopt the proposed innovation on a general, nonexperimental basis; to adopt it on an experimental basis; or to forgo it entirely. Experimentation should be considered only when a lack of particular knowledge precludes making any satisfactory choice between the innovation and the status quo.

Third, there must be no other practical means to resolve uncertainties about the effectiveness of the proposed innovation. If essential information can be obtained satisfactorily through simulation or other forms of research that do not directly affect the operation of the justice system, considerations of ethics, and perhaps of practicality and economy as well, militate against a program experiment.

Fourth, the experiment must seriously be intended to inform a future choice between retaining the status quo or implementing the innovation. Thus, the information sought must pertain directly to the value of the proposed innovation. Program experimentation

should not be considered where fiscal, political, or other constraints are likely to preclude adopting the innovation on a general basis.

A last threshold condition may confine experimentation in the justice system to a narrower area than in science and medicine, where research is generally conducted on a wholly voluntary basis. Reasonable risks that subjects knowingly accept in these fields can be justified by the general scientific, medical, or social value of the information that the research may yield. As long as subjects are competent, adequately informed of the risks involved, and able to exercise their judgment freely, consent serves as a powerful safeguard against unwarranted experimentation.

Experiments within the justice system, however, almost always involve the imposition of some mandatory element upon the experimental subjects, because of the the mandatory character of the justice system itself. If the innovation would not be voluntary when generally implemented, an experiment involving consent might not be adequate to predict the consequences of the mandatory program. In the justice system, then, informed consent is often not available to serve the protective purpose it does in other fields.[6]

That there are threshold conditions for undertaking a program experiment in the justice system should not imply a general presumption against experimentation. Existing procedures or proposed innovations may also compromise individual interests or place burdens on the public. A decision to innovate without experimentation or even to retain the status quo requires the same careful consideration of practical, economic, and ethical consequences that is required for a decision to experiment. But with few exceptions, it will only be when neither of these alternatives is acceptable that a program experiment may be justified.

Consider the hypothetical example of a court that requires the parties in all civil cases to participate in a pretrial settlement conference conducted by a designated judge. Suppose that the judges of the court have come to question the value of requiring this conference. They suspect that the conference rarely produces positive results unless at least one of the parties actually desires to participate. It has been suggested that the conference be conducted only when at least one party requests it. The judges hope this procedure will save time and money that is now wasted by counsel, judges, and parties in unproductive, obligatory conferences. Yet there is

6. There are, of course, some types of medical and scientific research that do raise problems similar to those of experiments within the justice system, such as research on children or on adults who are incompetent to make the relevant decisions. Also, some experiments within the justice system can be conducted on a wholly voluntary basis. So the contrast between medical or scientific research and experiments within the justice system ought not to be drawn too sharply. Nonetheless, it would be misleading to ignore the general difference.

some uncertainty whether the consequences of this change will all be beneficial. Some obligatory conferences may have actually succeeded in achieving fair settlements and saved parties the expense of trial, although counsel had been unwilling to negotiate and would otherwise have forced the matter to trial. Assume further that the court is unable to find any other jurisdiction that has compared the two procedures, and there does not appear to be any satisfactory means to resolve their uncertainties other than some form of experimental test.

In this situation, the threshold conditions are met. There is doubt about the effectiveness of the existing mandatory conference procedure. There is uncertainty about the value of the proposed innovation: it may be better in some ways but worse in others than the existing procedure. There appears to be no satisfactory way to resolve the uncertainties without an experiment. Experimentation should therefore be considered.

Meeting these threshold conditions, however, does not demonstrate that any particular program experiment will be justified on ethical grounds. Decisions about the justification of particular experiments are the subject of Chapter VI. Before reaching this decision, one must first consider two additional factors: the potential of particular experimental designs to resolve uncertainties effectively, which is the subject of Chapter III; and the harms or risks associated with conducting an experiment, which are examined in Chapters IV and V.

CHAPTER III. EXPERIMENTAL DESIGN

> It may be accepted as a maxim that a poorly or improperly designed study involving human subjects . . . is by definition unethical. Moreover, when a study is itself scientifically invalid, all other ethical considerations become irrelevant . . . . A worthless study cannot possibly benefit anyone, least of all the experimental subject himself. Any risk to the subject, however small, cannot be justified. In essence, the scientific validity of a study on human beings is in itself an ethical principle.[7]

David Rutstein's observation regarding medical research applies equally to experimentation within the justice system. Experimenting without a clear understanding of the questions to be addressed and the means for discovering useful answers not only is likely to be wasteful and seriously misleading, but also violates a basic ethical requirement not to expose people to needless harm.

Effective design requires an understanding of what experiments can and cannot accomplish, and of how different aspects of experimental design may influence the results. This chapter presents an overview of the theory and methods of experimental design.[8] In doing so, it draws upon expertise from the sciences, where refinement of experimental methods has produced techniques capable of yielding highly certain answers to questions about cause-and-effect relationships.

A program experiment, in whatever field, seeks to discover whether the program produces intended consequences while avoiding unintended ones. Experiments are designed to test the cause-and-effect relationship implied by such questions as "What level of fertilization produces the optimal yield in production of sugar beets?" or "Does simple mastectomy afford the same rate of survival as radical mastectomy?" or "Does the 'new math' result in greater comprehension of mathematics than traditional teaching methods?" or "Does court-annexed arbitration reduce the incidence of civil trials?"

7. D. Rutstein, *The Ethical Design of Human Experiments*, in Experimentation with Human Subjects 384 (P. Freund ed., 1970).

8. Appendix B presents a more thorough discussion, useful for the administrator planning an experiment, but not necessary to an understanding of this report.

"Experimental methods" refer to scientific techniques devised to achieve reliable and valid conclusions about particular kinds of cause-and-effect relationships. An experimental research method is needed to answer the question "Does this particular halfway house program succeed in reducing narcotics use by parolees with a history of narcotics addiction?" Experimental methods are distinguished from descriptive research, which might be employed to answer the question "What percentage of parolees completing the halfway house program refrain from subsequent narcotics use?" The first question seeks a comparison between drug use among parolees exposed to the halfway house program and the drug use they would have experienced in the absence of the program. The latter question simply asks for a description of drug use among parolees exposed to the program, but does not inquire about a causal relationship. A finding that 50 percent of the parolees completing the program refrain from subsequent narcotics use would not demonstrate that the program causes a reduction in narcotics use. Knowledge of some additional information (for example, that 80 percent of all previously addicted prisoners return to drug use upon parole) is essential to any inference that the program reduces addiction.

An inference about a causal relationship thus rests upon a comparison between what occurred with the program in operation and what would have occurred without it. The confidence one may rightly place in such an inference depends on the validity of the underlying comparison. If the halfway house program is applied only to selected parolees judged particularly suitable for halfway house treatment, participants may be atypical of parolees in general. In that case, the comparison of 80 percent addiction for typical parolees with 50 percent for the select group of program participants would be highly suspect. It could well be that the program does not reduce addiction, but rather that it selects participants who are especially likely to overcome addiction with or without the program.

A. Creating or Identifying Groups to Be Compared

It is sometimes possible to study the effects of different treatments applied to the same subjects, for instance, a series of different analgesics given to a single group of chronic pain sufferers. The nature of the justice system, however, usually makes such comparisons impossible—one cannot, for example, conduct two different types of trials for the same case. Instead, it is almost always necessary to compare results of the innovative treatment applied to one group of subjects with results of the present or alternative treatment applied to some other group. Such comparisons are valuable only if the groups are truly comparable.[9]

The comparison can take a number of forms. Results of an experimental program can be compared with results obtained from a randomly selected group that does not participate in the program; or they can be compared with results obtained from some other group chosen for its similarity to the program group; or they can be compared with results obtained before the program was put into effect. These approaches, or designs, are discussed briefly in the following sections, with an evaluation of the clarity of inference that each may allow.

1. Randomized Designs

Randomized experimental designs are an especially useful starting point because they best illustrate both the methodological strengths and the ethical problems of rigorous forms of experimentation.[10] In its simplest form, a randomized design requires that potential program participants be divided randomly—that is, by lottery—into two groups: an experimental group to which the experimental program or treatment is applied, and a control group, which receives the status quo treatment or some other program with which the experimental program is to be compared. The characteristics or actions of participants that the program is expected to affect are then monitored. If differences between the groups are sufficiently clear in statistical terms,[11] those differences can be understood as effects caused by the differences in treatment.

In any experiment, differences between the groups exposed to different treatments can stem from any of three sources: (1) differences in the treatment or experience of the groups, (2) preexisting systematic differences between the groups, or (3) differences between the groups that arise when characteristics of the subjects happen, purely by chance, to be distributed unequally between the groups. All experimentation seeks to eliminate the second and

9. Comparisons of noncomparable groups can be valuable if sophisticated methods of analysis mentioned in Appendix B can be employed (see pages 110-112). But these methods are applicable only under conditions in which the differences between the groups and the causal influence of those differences are very well understood. Those conditions can rarely be met in program experiments.

10. This report's emphasis on randomized designs should not suggest that randomization is always the preferred mode for program experiments. But its combination of methodological advantage and ethical disadvantage often makes it the most challenging example in the ethical analyses presented in subsequent chapters.

11. Whether a difference is clear in statistical terms will depend on the size of the groups, the magnitude of the difference between groups, and the extent of usual variation in the matter observed. Known technically as "statistical significance," the concept is discussed in more detail in Appendix B (see page 93).

third explanations for differences, leaving the first as the sole basis for causal inference.

What distinguishes the randomized design from others is the random assignment procedure, which tends to assure that any initial differences in the characteristics of individual participants are distributed equally between the two groups. Since assignment of an individual to one group or the other is without regard to any characteristic of the individual, the average characteristics of the groups will not differ in any systematic way. The second potential explanation for differences between groups is simply inapposite. Any initial differences between the groups must be nonsystematic—explainable only by the laws of probability. If the behaviors or outcomes of the groups are subsequently found to differ to an extent that cannot feasibly be attributed to chance, then the difference can only be accounted for by differences in the treatment or experience of the groups occurring subsequent to their random creation.

Statistical techniques can determine the mathematical probability that a particular difference could have arisen from a chance imbalance in the groups, which often enables us to dismiss the third explanation as improbable. Thus, random assignment and statistical methods together can narrow the potential explanations for observed differences, leaving an unambiguous inference that differences between the groups were caused by differences in treatment or experience following randomization.

In contrast, groups selected without randomization will always differ in some systematic way other than exposure to the experimental program. Statistical techniques can eliminate chance as a feasible explanation for differences, and thus narrow the explanations for difference to two. But without randomization there are no certain methods for determining that observed differences between groups are not related to the preexisting, systematic difference. An experimental comparison between systematically different groups will produce ambiguous implications whenever the systematic difference affords a plausible explanation for apparent effects of the experimental program.

If a randomized experiment were used to evaluate the halfway house program mentioned previously, potential participants would be assigned randomly to either an experimental group placed in the halfway house or a control group receiving the status quo treatment. Differences between those two groups in subsequent narcotics use would reveal effects of the program.

Randomized experiments may also employ random assignment of groups of persons, or of institutions such as courts or prisons; they are not limited to random selection at the individual level. For example, if a number of district courts were involved in an experiment, it would be possible to assign entire districts randomly. Random assignment of each of twenty districts to either the experimental or the control group would allow inferences about the program's effects on both the functioning of the districts and the behavior of individuals within the districts.

Although randomization eliminates preexisting systematic differences, the use of random selection to determine who receives the experimental treatment may itself cause problems if participants know of its existence and therefore behave differently. And it is sometimes very difficult to ensure that the only difference in subsequent treatment of randomly selected groups is the difference that was intended for purposes of comparison. For instance, participants in an experimental halfway house program may be treated differently by the police because of their special status, and that difference may contaminate the measurement of recidivist behavior in the experimental group. Nonetheless, for many program experiments, randomization permits more credible inferences about effects of the experimental treatment than does any alternative design. The strengths of the randomized experiment are perhaps best appreciated when contrasted with the potential weaknesses of alternative strategies.

2. Comparison Group Designs

It is often possible to locate two existing groups that appear to be similar in ways relevant to the program to be tested, but that are exposed to different programs, and then to compare the two groups in order to draw inferences about the program's effects. Such "comparison group" designs permit evaluation of program effects by using differences in treatment that occur naturally, or by manufacturing such differences intentionally but not randomly. For example, if participation in the halfway house program is voluntary, the program's effectiveness might be tested by comparing narcotics use of the volunteer-participant group to that of the nonvolunteer-nonparticipant group.

Problems arise in this type of comparison because differences observed between the groups can often be explained by potential causes other than the experimental program—that is, by rival hypotheses. If volunteers choose the program because they are more interested in avoiding narcotics use, they are likely to experience less subsequent narcotics use than the nonvolunteers, whether or not they participate in the program. This possibility reduces the credibility of experimental results that suggest the program is ef-

fective in reducing narcotics use. Randomly assigning participants to the two groups, by contrast, would more definitively rule out any such alternative explanation based on preexisting dissimilarities between the groups.

In some instances, however, comparison group designs can produce very credible results. If preexisting differences between the groups could not reasonably account for differences in outcome as substantial as those anticipated from the experimental program, the design's potential ambiguity may be insufficient to warrant any skepticism. For instance, if there were evidence that two jurisdictions, one with a halfway house program and one without, had substantially similar patterns of narcotics use, and if it were possible to identify persons in the jurisdiction without the halfway house program who would have been placed in the program if they had been in the other jurisdiction, a comparison group design could be a useful research method. Yet in many cases it is difficult to ensure that the comparison group is sufficiently similar to the experimental group, and the validity of any inferences about the program's effects will therefore be uncertain.

3. Before-After Designs

"Before-after" designs permit comparisons using the same category or population of subjects at different periods of time.[12] The comparison is between the results of the status quo, obtained before the experimental program was instituted, and the results obtained thereafter. Narcotics use by parolees who participate in a halfway house program, for instance, may be compared to the narcotics use of their counterparts paroled before the halfway house program was established.

A common problem with this design is determining which individuals in the past comprised the population for which the program is designed. That problem may make it impossible to produce a reliable comparison with the "after-the-program" group. If the halfway house program is voluntary, it might be impossible to determine which of the "before" parolees would have volunteered for the program if it had been offered.

Before-after designs are also subject to some of the problems of comparison group designs. Some relevant dissimilarity other than exposure to the experimental program could cause differences thought to be effects of the program. With the passage of time, many changes occur in a population and its environment; the ability to exclude possible effects of such changes determines the credibility of inferences derived from before-after comparisons. One needs strong evidence of such ability in order to rely on before-after designs.

12. See, for example, the experiment described in Chandler v. Florida, 49 U.S.L.W. 4141 (1981).

Despite these difficulties, before-after designs have their place in experimental evaluation. When characteristics thought to be affected by the experimental program are stable over time, and when the appropriate "before-the-program" group can be identified, a before-after comparison can warrant confident inferences about the program's effects. Such research designs are suspect, however, when time-related changes occur frequently in the population in question or when the effects of the experimental program are likely to be subtle.

Consider an experimental program designed to increase the rate of pretrial settlement in some class of civil cases for which the settlement rate has historically fluctuated between 80 and 90 percent. If the settlement rate for cases litigated after the program is implemented does not substantially exceed 90 percent or fall substantially below 80 percent, the results will be ambiguous. Suppose the "after" rate is 85 percent. That could be an improvement over what would otherwise have been a normal fluctuation to as low as 80 percent, or it could just as plausibly be a deterioration from what would otherwise have been 90 percent. If the question is whether a proposed innovation produces significant but subtle improvements in similarly unstable conditions, before-after analysis is usually not adequate. If, on the other hand, the consequences of a program will be dramatic if they occur at all, or if they must be dramatic in order to warrant the costs or harms associated with the program, a simple before-after design may well be quite adequate. Thus, if the program just mentioned had to achieve a settlement rate of more than 95 percent to be considered worthwhile, a before-after design could probably provide the needed information.

Before-after designs are often chosen by default because little forethought is given to experimental design before an innovative program is instituted. When the opportunity to construct a randomized experiment or to identify an appropriate comparison group—either of which might be more appropriate—is lost, before-after analysis applied to routinely collected statistics should be greeted with considerable skepticism. Routine statistics are often inadequate, whether as measures of the factors the innovation is designed to affect or as bases for identifying the relevant "before" group.

An especially pernicious difficulty of such after-the-fact analyses is that innovative programs are often implemented in response to a sudden exacerbation of a problem, even though the change may

simply be an unusual variation in a naturally fluctuating pattern, as in the rate of pretrial settlements. The program could then be predestined to appear effective if the factors contributing to the problem would in any case have returned at some point to their historic level. Moreover, avoiding after-the-fact analysis requires prior attention to the goals of both the experiment and the experimental program. This attention can improve the value of information derived from the experiment as well as the quality of the future program.

B. Relevance and Comparability of Measurements

The validity of inference from experimental results will depend not only on the comparability of groups, but also on the means used to measure results and the relevance of such measurements to the questions at issue. Regardless of the choice of experimental design to study the halfway house program, the narcotics use of all subjects (those participating in the halfway house as well as those serving as controls) would have to be monitored accurately. A method for measuring narcotics use would be required—one that is both reasonably accurate and applicable to both groups.

A number of such methods will often be available. In the case of narcotics use, for example, perhaps the most accurate would be provided by weekly urinalysis tests of all subjects. Records of arrest or conviction for narcotics offenses would be less accurate but possibly satisfactory, if one could be confident that the incidence of arrests or convictions reflects actual narcotics use with reasonable accuracy. Even less suitable would be some very indirect measure, such as the frequency of subject participation in a voluntary program of therapy for ex-addicts.

No special techniques exist to overcome the problems of assuring either the relevance or comparability of measurements: the nature of the groups to be compared often precludes the use of any adequate yardstick. In these instances the only solution lies in choosing other groups, to which a satisfactory yardstick can be applied. The most serious problem with employing a before-after design can be the virtual impossibility of applying a satisfactory yardstick to both groups, even though the groups themselves may be quite adequately comparable. Similarly, the advantage of a randomized experiment may be not only in the comparability of groups produced by randomization, but also in the creation of contemporary and equally accessible groups to which a satisfactory yardstick can be applied. Thus, a randomized design might be superior for the halfway house experiment simply because it would allow collection of weekly urinalysis data from both experimental and control groups.

C. Comparability of the Experimental Treatment to Its Future Nonexperimental Application

Regardless of the apparent reliability of the experimental design and yardstick employed, credibility of results can be undermined by factors that distort the behavior of participants or the experimental program itself.

Social scientists have demonstrated that people who know they are being studied often do not behave as they would without that knowledge. Participants who know that a program is experimental rather than routine may behave differently than they would if the same program were established on a nonexperimental basis. Consequently, the experiment might seem to show that the program was ineffective, when in fact it simply showed that the program was ineffective when implemented in the experimental context, although a nonexperimental program might work quite well. Similarly, experimental subjects who perceive an innovation as "new and better" might assess their experience more favorably than they would if the program were thought to be routine. From a purely methodological point of view, the obvious solution is to conceal the experimental aspect of the program from the participants, but ethical constraints may preclude that choice.

Finally, credibility of results requires that the experimental program has, in fact, been implemented and conducted in the manner intended. Even when rigorously designed, an experiment will produce unreliable results if the experimenters do not have a clear understanding of the program they are testing. For example, it is of little use to find that a rule calling for pretrial conferences in some class of cases does not reduce the incidence of trial, unless we also know the nature of the pretrial conferences actually conducted and the extent of adherence to the rule. The rule's failure to decrease the incidence of trials may be attributed to failure of the concept or to failure in its implementation. An effective experimental design must include a plan for thoroughly describing the implementation and operation of the experimental program.

CHAPTER IV. BASIC ETHICAL CONSIDERATIONS

The preceding chapter suggested that experiments involving random assignment of subjects to treatments, use of reliable but intrusive "yardsticks," and concealment of experimental purposes or actions can produce very accurate, unambiguous assessments of an innovation. But these features of a properly designed experiment conflict with ethical principles favoring equal treatment, individual autonomy, respect for privacy, and candor. Encroachments upon these values represent the ethical price to be paid for the benefits of experimentation. Even a temporary encroachment is justifiable only if narrowly confined and if likely to provide an important contribution to our system of justice. The question in each case is whether the benefits exact too high a price. And there are, of course, types of encroachment that are unacceptable in any circumstance.

Ordinarily, evaluating the ethical strengths and weaknesses of a proposed experiment must involve a careful balancing of the anticipated harms against the benefits expected from the experiment. Of course, an innovative program itself may generate harms and benefits, but the techniques for their evaluation are beyond the scope of this report, which is concerned only with the ethical questions involved in a decision to experiment. The proposed program's anticipated results should have been adjudged socially desirable before the difficult ethical issues involved in experimentation are confronted.

Before turning to a detailed analysis of the harms and benefits involved in the decision whether to experiment (see Chapters V and VI), we will consider here some general ethical principles that provide the framework for the later discussion. These general principles provide important ethical guidance to those who must decide whether, and how, to experiment with individuals and institutions in the justice system.[13]

---

13. Our formulation of principles owes much to the work of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, to which the committee gratefully acknowledges its indebtedness. (The Belmont Report, *supra* note 3.) But this committee's interpretation of principles is clearly not the same as the commission's, because of the special place these principles al-

A. Equality of Treatment

Equal treatment is a principle of fairness requiring that individuals who are similar in relevant ways be treated similarly. The ethical principle of equal treatment enjoys special status in our legal system, and is therefore of special relevance in the design and conduct of program experiments.

Program experiments usually entail introducing an innovation on a limited basis; not all persons who are similar in ways relevant to the innovation are afforded its benefits or exposed to its harms. Sometimes the rationale for limited application is economy. A pilot program may be tested in only a few locations so that implementation costs may be saved if the innovation proves unsuccessful. Disparity may also be created to help ensure that the experiment will provide valid and reliable information, in accordance with appropriate standards for the design of experiments.

Whether it is chosen for purposes of economy or credibility, disparate application of an experimental program presumptively conflicts with the principle of equal treatment and harms individual interests. Experimental disparity emerges in its sharpest form when a program involving harm to subjects is mandatorily imposed on randomly chosen individuals. Disparity, whatever its purpose, is a pervasive and serious ethical problem in program experimentation.[14] A dominant concern of this report is to analyze experimental disparity, the burden of justification associated with it, and the countervailing benefits that may or may not meet that burden.[15]

The principle of equal treatment requires that the harm or risk associated with program experimentation be allocated equitably. A particular class of persons should not suffer an undue share of harms or risks. Programs that are ultimately intended for application to all civil cases or all prison inmates ought not to be tested initially on particular groups of litigants or offenders because those

ready occupy within our system of justice. We are not importing principles from outside the system to evaluate the design and conduct of experiments; rather, we are applying to the relatively new field of program experimentation existing principles to which our system of justice is already deeply committed.

14. Of course the ethical problem is not unique to experimentation. In a variety of contexts, randomness as a basis of classification is recognized as not ultimately incompatible with the norm of equal treatment and indeed is accepted as an ethically desirable procedure: *e.g.*, the selection of jury panels, the assignment of cases to judges, or the order of call in compulsory military service.

15. This report's use of the terms "harm" and "burden of justification" should be spelled out. Experimental practices that conflict with the principle of equal treatment or the principle of respect for persons are considered as harms, in order to emphasize the special place these principles occupy in our legal system. These harms range from modest to severe, and carry corresponding burdens of justification. "Burden of justification" simply refers to the weight of benefits necessary to justify the harm.

groups are, for example, too powerless or passive to contest being singled out for experimental purposes.

B. Respect for the Person

The principle of respect for persons favors actions that respect the autonomy, integrity, privacy, and dignity of individuals. Treating rational adults in accord with this principle means respecting their judgments regarding what is to their benefit and in their interest. Within broad limits, it should be presumed that individuals are entitled to make their own decisions on matters affecting their lives. There are ways in which government is absolutely prohibited from invading individual autonomy, but it is important to emphasize that invasions not thus prohibited must be supported by properly delegated authority and adequate justification.

Respect for persons requires that, whenever possible, consistent with experimental objectives, experiments should be conducted only with the participants' fully voluntary and informed consent. This means that there is harm in compelling a subject to participate in an experimental program, in excluding any individual for whom the program is intended, in compelling a subject to divulge information, or in allowing a subject to be misled about the nature and purpose of the experiment.

Concerning children, mentally incompetent adults, or others incapable of exercising autonomy, respect for persons requires providing adequate representation and protection of their rights and interests. It is crucial that their interests be represented independently and competently in any decision about their participation in a program experiment. Children or the mentally incompetent may sometimes be subject to experiments mandatorily, but in no case should the experimenter be allowed to decide what is in the best interests of such persons.

The justice system frequently restricts autonomy in various ways: by imprisoning, by compelling obedience to judgments for damages, and, less drastically, by imposing rules of court procedure, establishing priorities for the use of law enforcement resources, and setting guidelines for parole decisions. Respect for persons requires that any additional mandatory requirement imposed for the purpose of experimentation carry its own burden of justification.

C. The Justification for Imposing Harm

Individuals should be exposed to harm or risk only when the expected benefit clearly outweighs the burden or harm. Alternative experimental designs can be evaluated by asking whether the greater benefit available from one alternative will clearly outweigh an associated increase in harm.

Even when the benefits clearly outweigh the risks of harm, however, no experimental method should be employed when a less harmful, reasonably available method can produce the information needed. If a less harmful alternative is likely to produce less adequate information, the more harmful alternative can be justified only by comparing the increased harm with the increased probable benefit.

Whether expected benefits clearly outweigh harm done to individuals will of course depend on the nature of the benefit as well as that of the harm. Benefit to persons or groups other than those harmed in the course of the experiment will carry much less justifying force than expected benefit to the individuals harmed. As the certainty and significance of harm to individual subjects increases, it will become correspondingly difficult to consider the harm to be clearly outweighed by benefits to others. Because, in general, those most likely to benefit from a program experiment are future members of the class of experimental subjects, while those most likely to be harmed are the actual subjects, the requirement that benefits clearly outweigh harms must be regarded as a stringent standard for responsible program experimentation.

D. Harms to the Public

Program experiments may harm the public in general, as well as individuals who participate in them. Harm to the public always requires justification, but need not require the same kind of justification that is demanded when individuals participating in experiments are harmed. Because the essential reason for program experimentation is the public's interest in informed decisions that facilitate the effective administration of justice, harm to the general public that may ensue from an experiment can generally be evaluated through a more direct weighing of costs against benefits. The financial cost of conducting an experiment, for instance, can be balanced against anticipated savings.

An experiment may have harmful or beneficial effects on various interests of the public. Among those effects are economic consequences (of the innovation or of maintaining the status quo, as well

as the costs of conducting the experiment) and any other potential consequence that could affect individuals simply because they are members of the society in which the experiment takes place. Indeed, the practice of disparate treatment might be regarded as a public harm if it would create an appearance of inequity that could undermine public faith in the justice system.

The public benefits likely to ensue from an experiment must be evaluated in light of the factors that initially led to consideration of program experimentation. Factors mentioned in Chapter II—inadequacy of the status quo, uncertainty about the effectiveness of the proposed innovation, and lack of alternative means for resolving those uncertainties—identify the information needed from an experiment and the importance of obtaining it. The benefit that might ensue from the results of an experiment is limited by the potential for improvement over the status quo, by the increased cost and potential adverse consequences of the innovation, and by the possibility that uncertainties could be resolved by other means, even if more costly, less certain, or less practical. The evaluation of benefits as justification for harms is explored more thoroughly in Chapter VI.

E. Applying General Principles

Although the principles we have identified provide a basic framework for determining whether and how a program experiment should be conducted, they also underscore the difficulty of the task. Considering a program experiment calls for a measurement of benefit, principally in the form of reliable information about program consequences, weighed or balanced against such harms as disparate treatment of similar persons and limitation of individual autonomy.

It is only metaphorically that one can speak of "weighing" or "balancing" such incommensurate factors. In easy cases, where an experiment offers great benefit and minimal harm, or great harm and minimal benefit, the metaphor provides a form for articulating the obvious conclusion that the experiment should or should not be performed. In difficult cases the suggested "weighing" will provide a procedure, but cannot provide a formula to guarantee correct answers. The best available answer will be a judgment made in good faith, and reasonable people will sometimes disagree. This does not mean, however, that ethical judgments about program experiments are always to be made on an isolated, *ad hoc* basis.

The context for judgments will include alternatives of adopting the innovative program without prior experiment or of simply

maintaining the status quo. Either alternative will present its own set of harms and benefits. These alternatives are frequently chosen in the administration of justice, and those choices provide some guidance on the mixtures of harms, risks, and potential for benefit that are normally thought acceptable. Such norms provide a useful standard for evaluating proposed experiments.

Examining analogous harms that are common practice in the administration of justice can also guide evaluation of the particular harms associated with experimental designs. Although random assignment of court cases to different procedures is not commonplace, there are relevant similar practices: for instance, the "pilot test" of an innovative procedure in one court among several, or variation in local rules and in the practices of individual judges. The existence of some forms of disparity obviously does not automatically justify other forms. But it is reasonable to compare forms of disparity in order to judge their relative harm. The acceptability of practices that are commonplace in nonexperimental settings is a guide for evaluating the same or similar practices undertaken for experimental purposes.

There will not be universal agreement about what should be included in each of the various classes of harm or benefit arising from program experiments. Does an experimental rule of civil procedure limit individual autonomy, so that it must be counted as a harm? Does an experiment involving the elimination of oral argument for some cases entail disparity that affects any substantial interest of litigants? Is routine, compulsory taking of blood samples from parolees in order to monitor narcotics use a harm of a type that cannot be justified?[16]

These are matters that general principles alone cannot satisfactorily resolve; they must be decided by judgment in the circumstances of particular proposed experiments. Accordingly, the committee has not tried to define the boundaries of either individual or societal interests that may be affected by experiments. We recognize, for example, that experiments may entail infringements of privacy, but we have not tried to define privacy interests or the limits of justifiable intrusions of privacy. Instead, we offer a framework in which the practices and consequences of experiments may first be recognized as harms or benefits and then be accounted for in deciding whether to undertake an experiment.

16. In using these illustrations, the committee takes no position on the merits of the issues. We simply recognize that they may generate responsible disagreements.

CHAPTER V. ANALYZING HARMS IN PROGRAM EXPERIMENTS

A. The Harm of Disparate Treatment

When a program experiment is undertaken, it should be with the expectation that the program being tested would apply to some definable class of persons if it were adopted on a general basis. Members of that class are the potential subjects. An experiment creates experimental disparity whenever the experimental program is applied to fewer than all potential subjects. An experiment to test an innovative rule of court creates experimental disparity if the rule is experimentally applied only to a particular class of cases with the expectation that it will be applied to all civil cases if it is shown to be effective. Similarly, experimental disparity occurs when a state establishes an experimental pretrial discovery program in a single community, with the expectation that the program will apply throughout the state should it prove successful. Randomized disparity imposed at the individual level within the same jurisdiction or geographical area raises this problem in acute form.

Experimental disparity can be distinguished from other, closely related kinds of disparity. Disparity results when all cases before a particular court are subject to a program that a neighboring court has not adopted. Such disparity demands attention and justification, but the questions it presents are more analogous to those arising from a decision to adopt a program on a general, nonexperimental basis. Similarly, a special procedure that applies only to those cases that constitute the bulk of a court's backlog may create disparity in respect to other, similar cases before that court. These types of disparity present issues that are not peculiar to experimentation. Resolving such issues depends on the acceptability of nonuniform treatment as an aspect of the ongoing administration of justice, a subject that this report does not specifically address.

Experimental disparity always creates some harm that must be justified. The degree of that harm and the burden of justification it carries will depend on six factors:

1. the significance of the interests affected;
2. the extent of the difference between treatments;
3. a comparison of the disparity with standard treatment or expectations;
4. the degree to which the disparity reflects differences in qualification of subjects;
5. whether the experimental treatment is harmful or beneficial; and
6. whether participation is mandatory or voluntary.

1. Significance of the Interests Affected

Experimental disparity may involve interests of varying levels of importance.[17] A disparity in trial procedures, for example, will generally be regarded as more significant than a similar disparity in pretrial procedures. An experiment substituting magistrates for judges in voir dire may present more troublesome disparity than one in which magistrates conduct pretrial conferences because voir dire is an element of trial, which is generally of greater concern to litigants than a pretrial conference.

2. Extent of the Difference Between Treatments

The extent of the difference between treatments—the magnitude of the disparity—will directly influence the associated burden of justification. In some instances measuring the difference is easy and quantitative. Disparate allocation of one-year and five-year terms of probation, for instance, is more severe than disparate allocation of one- and two-year terms. Rarely are matters so simple.

Consider, for example, disparate application of an experimental program requiring juveniles who commit offenses against property to make restitution. The class of potential subjects comprises juveniles who, under the status quo, would be sentenced to a short term of incarceration. Assume that incarceration involves a very brief period of detention, while restitution requires weekly obligations for an extended period. Assessing the difference between treatments—and establishing the consequent burden of justification—will be difficult. The difficulty lies in comparing the severity of two dissimilar harms.

17. Although "experimental disparity" usually refers to the difference between experimental and status quo treatments, the expression encompasses more complex situations, such as when more than one experimental treatment is tested (*e.g.*, two alternative types of pretrial conference, tested against each other or against the status quo treatment).

The full range of possible harms should always be considered carefully. For instance, if some of the juveniles sentenced to incarceration are likely to suffer psychological or physical harm while they are institutionalized, the difference between restitution and incarceration may emerge as a substantial disparity. If some of the juveniles assigned to the restitution program need psychotherapy that is available only through incarceration, the harm of the disparity may be quite serious. If risks of this kind cannot be predicted or avoided even under the status quo, however, then they present no additional problem in an experiment involving disparate application of the restitution program.

3. Comparison of the Disparity with Standard Treatment or Expectations

The difficulties arising from experimental disparity depend to a significant extent on a comparison with the status quo. Under the status quo, all persons might receive identical treatment. Or, they might receive different treatments according to procedures other than those to be employed in the course of an experiment. Conventional procedures for allocating different treatments may be of two kinds. One involves the discretionary assignment of different treatments on an individualized basis according to the need, merit, or desert of each individual. Another procedure involves explicit rules under which separate categories of persons or cases are given different treatments. Each should be contrasted with the disparity involved in the experiment in order to assess the difference between the status quo and the experimental disparity.

a. Experimental Disparity Compared with Individualized Treatment

The individualization of decisions about treatment of individuals reflects an important and accepted value in certain areas of the legal system—in sentencing, parole decisions, and decisions in which judges or other officers have broad discretion, such as the use of a court-appointed expert witness or the prosecution of a particular offense. Individualized judgments further the principle of respect for persons because recognition of the uniqueness of individuals and their circumstances honors the concept of human dignity that is at the heart of the principle. Moreover, individualized judgments accord with the principle of equal treatment, which connotes identical treatment only when there are no identifiable dif-

ferences among individuals that would suggest differences in treatment. When there are differences in the "qualifications" of individuals—differences in need, merit, or desert—equal treatment calls for differentiation according to qualification. When differences in individuals' qualifications are important and too subtle or complex for classification according to explicit rules, individualized judgments may still be achieved by experienced and conscientious decision makers applying implicit standards.

Random assignment to treatments, when substituted for individual judgments, conflicts with the principle of equal treatment. Random selection is by definition blind to differences in individual characteristics. Although program experiments of a nonrandomized nature may also be blind to relevant differences among individuals (as in a comparison group experiment where the disparity between groups is solely on geographic grounds), the contrast between randomized and individualized treatment is both unavoidable and usually stark.

A randomized experiment using treatments that would ordinarily be assigned according to individualized judgments carries a heavy burden of justification. Our system of justice attaches great value to the good-faith attempt to tailor treatments to individual circumstances. Assigning treatments according to the demands of an experiment means suspending that attempt. So a choice to forgo individualization, whether for random assignment or another process, must carry a substantial burden of justification even when there is uncertainty about the value of results achieved by actual individualized judgments. The good-faith attempt to individualize is itself valued, independently from the value of the results. If it were believed, for example, that existing disparities in sentencing are so great that the results amount to randomness, that alone would not justify allocating sentences on an intentionally random basis.

But it should be recognized that it is possible to justify suspending individualized judgments for experimental purposes, even, for example, to justify random assignment of sentences to offenders. Strong justifications can arise when the status quo is believed to produce harmful results and the proposed experiment is likely to produce important improvements in the results of future individualized judgments. Questions about justification are addressed in detail in Chapter VI.

b. Experimental Disparity Compared with Identical Treatment

Identical treatment often contains an element of arbitrariness, rather than a judgment that the treatment given to all is ideally suited to all. The kind of ethical difficulty associated with disparity that replaces individualized treatment will not necessarily occur

when disparity replaces identical treatment of all individuals in a category. But such disparity may offend expectations of identical treatment or create perceptions of injustice. A court may require pretrial conferences in all civil cases, for example, not because conferences are warranted in all cases, but because they are useful in most cases, and no satisfactory way has been found to assign them on a limited basis. If the court were to experiment by assigning only some cases to a promising alternative mechanism—informal, nonbinding arbitration, for example—it would not thereby abandon any special standard of care in assigning cases to treatments. Nonetheless, such an experiment might well violate important expectations of litigants and create perceptions of unfairness.

If the disparate treatment offends individuals' shared expectations of identical treatment, the harm of disparity may be aggravated. Suppose the experiment mentioned in the preceding paragraph involved random assignment of all civil cases before a single court—either to arbitration or to a pretrial conference. That might exacerbate the basic harm of the disparity by offending reasonable expectations that all litigants before the court will be treated identically in such matters. Contrast with that an alternative randomized experiment involving several courts, in which some courts would continue to employ pretrial conferences, while others, randomly chosen, would employ arbitration in all cases. In both experiments the determination of treatment would be based on a random decision, giving rise to the same basic disparity. The difference lies in the extent to which the experiments may offend reasonable expectations of identical treatment. This can be an important factor in choosing between the two types of experiment.

Further, disparate treatment may erode a valuable sense of commonality in a particular community. If the disparity an experiment produces is viewed as involving unfair privilege for some or unfair deprivation for others, the experiment may generate envy or resentment. The harm of the underlying disparity is aggravated by introducing an experiment where there are prior shared expectations or communal bonds. Assessing the risk of harm from an experiment therefore requires evaluating the presence or absence and the strength or weakness of such expectations or bonds in the affected population.

4. Whether Disparity Reflects Differences in Qualification of Subjects

Disparity occurs whenever an experimental program is applied to less than all those for whom the program is ultimately designed.

Yet among the class of persons for whom the program is designed, some may be more qualified to participate than others. The harm of disparity will generally be mitigated if differences in treatment of individuals accord with those individuals' differing qualifications.

Consider an experimental halfway house program designed to reduce recidivism among parolees. The program will entail significantly greater restrictions on the liberty of participants than does the status quo of straight parole, and it will be imposed mandatorily. The program will therefore subject participants to significant harm, which may or may not prove justified by the benefit of reduced recidivism. Because the program's purpose is to reduce recidivism, parolees especially prone to recidivism might be viewed as more qualified than those less prone. If the likelihood of recidivism can be predicted for various types of parolees, the harm of the disparity would be mitigated by imposing the experimental program only upon those most likely to return to crime.

The ethical advantages of such a procedure, however, must be contrasted with the disadvantages. A test on recidivism-prone parolees might show no positive results, although the same program applied to less recidivism-prone parolees might have very favorable results. Yet application of the program without regard to differences in qualification would aggravate the basic disparity. The harm of the disparity must therefore be balanced against the risk that the experiment will yield inadequate information.

Reasonable people may disagree about how qualification should be measured in particular cases. In the case just illustrated, one might argue that those most qualified are those most likely to benefit from the program (*i.e.,* parolees with a moderate chance of recidivism), and not those most in need of the benefit (*i.e.,* those with a high chance of recidivism). Nonetheless, whatever the accepted yardstick of qualification, the harm of disparity will be mitigated to the extent that the disparity accords with differences in qualification.

5. Whether the Experimental Treatment Is Harmful or Beneficial to Subjects

Whether the experimental treatment is harmful or beneficial to subjects will also affect the burden of justification associated with experimental disparity. Disparate imposition of harm demands greater justification than disparate imposition of benefit. But harm and benefit are relative, rather than absolute, concepts; a particular treatment is harmful or beneficial only in comparison to some

alternative treatment. Characterizing disparity as disparate harm or disparate benefit therefore depends on identifying the "relevant alternative" to which the treatment of subjects should be compared.

If an offender is committed to a halfway house under conditions of restricted liberty, for example, that treatment is harmful compared to conventional parole, but probably beneficial compared to continued incarceration. If an experiment involves disparate treatment of offenders, some of whom receive straight parole and others halfway house treatment, do those in the halfway house suffer a disparate imposition of harm or are those on straight parole afforded a disparate benefit? The relevant comparison is not necessarily the treatment that subjects would have received prior to the experiment; rather, it is the treatment they would have received in the absence of the experiment. That is, between the two alternatives to experimentation—innovating without experiment or forgoing the innovation and retaining the status quo—which would the administrator choose if the experiment were not undertaken? If all subjects would have received straight parole, then those experimentally assigned to the halfway house are disparately harmed. But if all would have been assigned to halfway houses, then the experiment creates a disparate benefit for those who receive straight parole.

The relevant alternative is often clear in light of the two contexts in which program experimentation is usually considered. In one, the innovation is potentially very costly, either financially or in light of potentially adverse consequences to subjects or the public. Given its cost and the uncertainty about its effectiveness, the innovation would not be undertaken on a general basis without prior experimentation that demonstrates its value. In this context the relevant alternative—the treatment subjects would receive if the experiment were not undertaken—is the existing, or status quo, treatment. In the second context, the innovation is relatively inexpensive and poses no serious risk of adverse consequences. If a choice had to be made between retaining the status quo and implementing the innovation on a general basis despite uncertainties about its effectiveness, the administrator would do the latter. In this context, the relevant alternative is the innovative treatment, and not the status quo.

There are contexts in which identifying the relevant alternative is more difficult. Sometimes experimentation is undertaken to devise improved means of choosing among existing programs that are ordinarily assigned according to individualized judgments.[18] In

18. An example of this kind of experiment is discussed at pages 61-64, *infra.*

that context, it might be impossible to decide what treatment any particular individual would have received in the absence of the experiment, and therefore impossible to characterize the impact of the experiment as harm or benefit.

6. Whether Participation Is Mandatory or Voluntary

Even when the relevant alternative is apparent, reasonable people may sometimes disagree about whether an experimental program is harmful or beneficial in comparison. Consider an experimental program for mandatory, nonbinding arbitration as a prerequisite to trial in civil actions; assume that the relevant comparison is to the status quo in which no such program exists. Some will view the arbitration program as harmful because parties subject to the arbitration procedure will incur costs associated with the arbitration hearing. But because nonbinding arbitration is intended as an alternative to the greater costs and complexity of trial, others may see the program as a valuable service.

There will be occasions when it will indeed be difficult to determine whether the disparate treatment harms or benefits individuals or groups. But, when an experimental program is mandatorily imposed, the principle of respect for persons requires that mandatory imposition itself be recognized as a harm. The likely benefit from the experiment must therefore be sufficient to outweigh the harm of mandatoriness as well as the harm of disparity. Such experiments can be justified, but the burden of justification will not be light.

Disagreements about the harmful or beneficial character of an experimental program often cannot be resolved without experimentation (*e.g.*, to ascertain whether arbitration results in a net increase or decrease in the expense of litigation). But the principle of respect for persons provides important guidance regarding how a program ought to be viewed in the face of uncertainty about effects. It is always preferable to allow individuals to choose between the experimental program and the status quo. When offered the choice, the individual assumes the responsibility of weighing harms against benefits.

Experimental disparity will pose fewer ethical problems if participation is voluntary. Random assignment or other disparity-producing designs can employ voluntary participation. One method is to allow any qualified subjects to participate in the experiment, provided the subject consents to be assigned to either the experimental program or the status quo. That is, subjects consent to disparate treatment. Another method is to allow only some of the po-

tential subjects to participate in the experimental program if they choose; the others are given the status quo treatment, without any choice. That is, subjects are disparately allocated the opportunity to consent. In either case consent will only mitigate the harm; it will not render an otherwise harmful disparity entirely innocuous.

Consent will be feasible only when a significantly large number of qualified individuals can be expected to view the experimental program as offering them potential benefit, or at least no significant harm. Moreover, an experiment using consent will be useful only if future policy decisions would be aided by information regarding the program's success when applied to volunteers. Obviously that condition will be satisfied when the program is intended for voluntary application; it will often not be satisfied when the program is intended for mandatory application.

The difference between consent to disparate treatment and disparate allocation of the opportunity to consent deserves inspection. Disparate allocation of the opportunity entails an obvious denial of benefit to those not afforded the opportunity. If the relevant alternative to disparate allocation of opportunity to consent is to offer the program to all potential subjects, then the disparity must be viewed as a disparate allocation of harm and must carry a commensurate burden of justification.

Consent to disparate treatment can also be problematic. If an experimental program is desired by potential subjects, the opportunity of assignment to it constitutes a benefit. But if an individual who desires the experimental program must consent to random assignment in order to obtain it, where it could be made available to all, then that consent cannot be regarded as fully voluntary, and thus may not significantly alter the harmful character of the disparity. Hence consent to disparate treatment may make less difference ethically than it seems to at first, and in some situations an experiment incorporating this approach will not be very distinguishable from one involving disparate allocation of the opportunity to consent.

An experiment employing consent to disparate treatment is nevertheless generally preferable to any other feasible basis for subject participation.[19] When assignment to treatments is random, consent to disparate treatment ensures that all potential subjects who

19. The least ethically troublesome basis for subject participation is one in which subjects may freely choose one treatment or the other, but groups created in this way will rarely be adequate for valid inferences of program effects. One could provide a third choice by inviting subjects to be assigned to one or the other treatment by the experimenters (they would freely consent to disparate treatment). Such a procedure, however, will usually not provide a sufficient number of subjects assigned to treatments by the experimenters, and those obtained may be extremely unrepresentative of the population of potential subjects.

desire the experimental treatment are given an equal opportunity to obtain it. Disparate allocation of the opportunity, in contrast, may result in the opportunity being given to some who do not wish it, and withheld from others who do. More important, an experiment employing consent to disparate treatment is less likely to provoke resentment of potential subjects who are denied the beneficial experimental program. Obtaining consent to disparate treatment necessarily involves candor: the participants have to be informed that the program is the subject of experiment, they must actively be recruited to participate in the experiment, and they have to be told that the treatment will be provided to some but denied to others in order to achieve the purposes of the experiment. Disparate allocation of the opportunity to elect an experimental program, however, may leave those who are denied the opportunity uninformed about why they are denied it, perhaps resulting in resentment and perceived injustice.

The opportunity to participate often cannot be extended to all individuals who might legitimately complain of disparate treatment. Consent will therefore rarely obviate the need to justify disparate treatment as an experimental harm. It is equally important, however, to recognize that the consent of individuals affected by an experiment is always ethically preferable to mandatory participation. Making participation in the experiment voluntary for some individuals, using either of the methods discussed, will significantly reduce the harm and burden of justification, although it will rarely remove them altogether.

B. Harms Other than Disparate Treatment

Three forms of harm other than disparate treatment require careful attention in decisions about experiments within the justice system. First, the use of persons as means toward an experimental goal. All experiments with human subjects involve such use to some extent. Second, the acquisition or use of information in ways that may compromise the privacy of subjects. Third, a lack of full candor with subjects about the nature of a program or the means employed in an experiment. Although these elements are sometimes necessary for obtaining reliable information about the effectiveness of programs, they must be recognized as harms that carry a burden of justification.

1. Using Persons as Means in Experimentation

Any research or experiment that involves human subjects uses those subjects as instruments of the research, as means to the end of obtaining information. This is most clear in laboratory research involving preliminary tests of new drugs, in which human subjects are used solely for the purpose of ascertaining the physiological effects of the drug; there is no purpose to benefit the individual subjects by administering the drug. In any experiment using random assignment of subjects to treatments, persons become means because the assignment disregards the needs and desires of individual subjects. Persons are also used as means when they are exposed to a novel program in a simple pilot project, because the purpose of the enterprise is not exclusively to further the subjects' interests, but also to obtain information that may prove useful to future policy decisions.

Using persons as means conflicts with the principle of respect for persons only when the individual subjects do not consent. When competent adult subjects are adequately informed of the nature, purposes, and risks associated with their participation, their consent obviates any potential offense to their interests that might otherwise arise from using them in the experiment. In program experiments that cannot involve consent, care must be taken to avoid unnecessary objectification of individuals as mere means to the ends of the experiment.

Such concern, however, must be assessed in light of the normal and accepted use of individuals as means in the administration of justice. Nearly every rule or program that is uniformly imposed on a class of persons ignores some particular circumstances of individuals in order to serve a larger group. Individuals are categorized as members of some class, rather than recognized as unique individuals. Thus, a rule requiring a pretrial conference in every civil suit may be imposed even though some cases will not benefit from it. The rule is imposed uniformly to reach all cases in which settlement might be obtained and trial avoided. Such common practices lessen, but do not eliminate, concern arising from the use of individuals as means toward an uncertain end.

2. Compromising the Privacy of Subjects

Reliable assessments of an experimental program's effectiveness may depend on information about individuals that would not ordinarily be available. Acquisition, use, or publication of such information may infringe privacy interests. The extent of harm or risk

and the consequent burden of justification will depend on the nature of the information obtained, the means used to obtain it, the use made of it, and the extent to which it is disclosed.

Two examples will illustrate several ways in which program experiments may compromise individual privacy. The first occurs in an experiment designed to assess a program to combat drug abuse by criminal offenders, which would require information about the incidence of drug use among both program participants and comparison subjects. Two potential measurements of that incidence are chemical analyses of blood samples and records of conviction for drug-related offenses. Although analysis of blood samples might provide a very reliable measure, obtaining the samples would entail a substantial intrusion of privacy, particularly if it were done without the subjects' consent—by force when necessary. Records of conviction, in contrast, would provide a much less reliable measure of drug use, along with a much reduced affront to privacy. Because convictions are matters of public record, there is no offense to privacy in merely obtaining the information. Nevertheless, the use or publication of such information may result in harm to subjects that would not otherwise befall them.

Information about an individual's history of drug use may, when known to others, result in significant harm by affecting social and vocational opportunities. The fact that such information is a matter of public record does not necessarily mean that it will be widely known. Experiments that increase awareness of such information present risks that must be justified by benefits to be gained. Moreover, the possibility of such harm requires utmost care in protecting the confidentiality of information or the anonymity of individuals.

The second example arises in an experimental program designed to reduce the expense of litigation in some class of civil cases. Because attorneys' fees are a major component of litigation expense, the experiment may require fee information for both participant and comparison cases. Mandatory disclosure could intrude upon the privacy of both litigants and attorneys. This concern might be ameliorated by using attorney hours as a surrogate measure of expense, or by effective methods for preserving anonymity.

Even after minimizing intrusions of privacy, an experiment will still almost always carry some risk of harm associated with potential disclosure of information. Exposing subjects to these risks carries an additional burden of justification.

a. Information Obtained Indirectly or Without Consent of the Subject

The affront to privacy is greatest when information is obtained by compulsion. Like mandatory application of an experimental program, mandatory disclosure of information must be presumed harmful to the individual's interests, carrying a substantial burden of justification. Voluntary provision of information is always preferred.

The experimenter's simple possession of information without subjects' consent may offend privacy. Even when sensitive information is obtained from some intermediate source, and only indirectly from the individual affected, the harm to privacy may be equivalent to that associated with mandatory disclosure.

Suppose an experiment with a drug abuse program uses information obtained during earlier, routine medical examinations of prisoners. The experimenter's acquisition and use of that information carries a burden of justification similar to that incurred if medical examinations were conducted for purposes of the experiment. If the examinations were performed without prisoners' consent in the first instance, then either the prisoners must consent to use of the information or the information must be regarded as compulsorily obtained by the experimenter and justified on that basis. If the examinations were originally conducted with the subjects' consent, any reasonable expectations or explicit guarantees of confidentiality must be respected. If the information is to be obtained in breach of those expectations or guarantees, it must be regarded as compulsorily obtained and justified on that basis. The problems associated with access to existing but confidential information can often be avoided by obtaining the information in anonymous form, through "file linkage" techniques (such as those discussed in Appendix B at pages 118-119). But even if anonymity is assured, some offense to privacy may remain. It must still be asked whether the use of information for purposes other than those originally envisioned will infringe the principle of respect for persons by making the subject an unwitting assistant in an endeavor he has not chosen to assist.

b. Information Obtained with the Consent of the Subject; Obligation to Protect Confidentiality

When information has been disclosed voluntarily, nonetheless harm may occur if confidentiality is not preserved. If dissemination of information would result in harm to the individual, the experimenter is obliged to minimize the risk of disclosure and to justify any disclosure that is required by the experiment. The obligation to

protect against such risks and to regard them as harms is particularly crucial when the individual at risk has been expressly assured, or may reasonably believe, that confidentiality will be preserved.

Obtaining sensitive information by exaggerated guarantees of confidentiality is a serious affront to the dignity of the subject—often more serious than obtaining the information by compulsion. It is doubtful that false or less-than-candid assurances of confidentiality could ever be justified in a justice system program experiment. The integrity of the system suffers when one of its officers promises confidentiality that is not or cannot be ensured. Great care must therefore be taken to avoid inflated assurances or expectations of confidentiality when soliciting sensitive information.

Privacy may be invaded even in using information that is a matter of public record and is therefore not legally protected against dissemination. This is a matter of concern not merely regarding the obvious, overt publication of names and embarrassing information, but also regarding more subtle public statements that may permit the inference of harmful information. An experimental program involving psychological diagnosis of prisoners in a specific prison, for instance, might reveal that all or nearly all of the inmates suffer from significant, chronic neurosis or psychosis. Publication of that information would plainly suggest that any person known to have been an inmate of the prison is psychologically disordered, resulting in clear harm to the inmates' future opportunities.

Although the risk of harmful disclosure of information must be justified in accordance with the benefit to be derived from the experiment, every effort should be made to minimize or eliminate such risk. In contrast to disparity or the need for information, disseminating sensitive information will rarely, if ever, further the purposes of an experiment. Such dissemination can often be avoided by effective methods for preserving anonymity or confidentiality. Such methods are discussed in Appendix B (see pages 118-121).

3. Deception: Compromising the Obligation of Candor

Behavior that a program is intended to alter may be affected not only by the program itself, but also by the subjects' knowledge that the program is experimental.[20] The subjects' reactions to "special" treatment may cause them to behave differently than if the program were established on a routine basis; that, in turn, may render

20. See page 23, *supra*, for a more thorough discussion of the methodological problems.

the results of the experiment ambiguous. It may therefore be important to ensure that subjects are not aware of the special or experimental nature of their treatment. From a purely methodological point of view, it might be desirable to avoid disclosing that subjects have been randomly assigned to different treatments, or that particular aspects of their behavior are being observed to determine how they are affected by the program, or that the program is in fact experimental and not routine. Thus, methodological rigor may require that subjects deliberately be deceived about the nature of the experiment, or that misconceptions they would naturally entertain be left uncorrected.

However, both active deception and allowing misconceptions to stand are severely at odds with a fundamental commitment to candor on the part of those who administer the law.[21] Scientists, too, have been concerned about the morality of deceiving research subjects.[22]

The obligation of candor in the administration of justice imposes a heavy burden of justification on any use of deception in program experiments. Deception on the part of those who administer justice poses one of the greatest threats to the integrity of our system of justice. This threat is most grave when the matter concealed through deception may itself appear to offend basic tenets of justice—concealment of disparity, for instance. Those who are deceived, and who are most likely to feel aggrieved by the concealed practice, would then be precluded from voicing their objections and from hearing the justifying arguments that might answer those objections. This may be contrasted to the harm of randomized disparity that does not involve deception. Although such disparity does harm individual interests, it does not necessarily undermine the manifest integrity of the justice system, provided the arguments advanced to justify the disparity are frankly disclosed. Such disclo-

21. This report uses the word "deception" to refer to failure to dispel misconceptions as well as overtly misleading statements or actions because candor in the administration of justice not only precludes overt deception, but also traditionally requires efforts to dispel misconceptions.

22. Codes of research ethics adopted or advanced in certain fields of science address this issue with great care, and some scientists believe that overt deception of subjects has no proper place in scientific research. Here again the distinction between experiments with subjects who participate voluntarily and experiments in which subjects participate without consent must be recognized. Any deception, whether overt or even unintentional, may undermine the "informed" nature of genuinely voluntary consent. So deception may be foreclosed in any experiment that depends on voluntary participation for justification either in law or in ethics. If subject participation is mandatory, deception is problematic not because it vitiates consent, but because deception is by itself an infringement of the principle of respect for persons and a threat to the integrity of the justice system. While deception need not therefore be prohibited absolutely, the burden of justification it must bear should be recognized.

sure conforms to the best traditions of our system of justice and reinforces its integrity. Deception and concealment, however, can be challenged by and defended to those deceived only after the fact.

Lack of full candor can only be justified by clear need to avoid ambiguity of experimental results that full candor would produce. Assertions that deception is necessary to avoid misleading results should be met with skepticism, and the decision to deceive should never be made without the concurrence of expert research methodologists. But the decision should not be delegated to researchers. The fundamental threat such practices pose to the integrity of the justice system requires that the decision to use them be made by a responsible justice system officer.

The burden of justification associated with deception depends on the significance of the matter concealed. If the matter concealed itself bears a substantial burden of justification, the deception must bear an even greater burden. Deception requires (1) that the concealment itself be indispensable to the validity of experimental results, and (2) that the burden of justification for the practice concealed not merely be met, but met by a clear and convincing margin.

Consider an experimental test of alternative types of citation for traffic offenses. Three types of citation are to be tested: a single written warning; a summons to appear in court; and the summons currently in use, which gives the driver a choice between appearing in court or paying a fine by mail. The purpose is to determine the relative effectiveness of the different types of citation in deterring future offenses by those receiving citations. In this experiment, the subjects would not be aware that they are being used as experimental subjects unless they were told so. The experimenters' failure to notify the subjects that they were involved in an experiment would be a form of deception that would have to be recognized and justified by benefits anticipated from the experiment.

In this example, the harm associated with the deception is relatively modest, and there are arguably substantial reasons supporting its justification. The matter concealed from participants is that an experiment is being conducted and that it involves a disparity in the rather modest harm or inconvenience associated with different forms of traffic citation. The burden of justification that such harms carry is relatively modest, so the burden associated with the deception is correspondingly modest. But to avoid the deception by informing subjects that the citation they receive is experimental and is issued in the course of a test of various types of citation may undermine the validity of the results. Any added deterrent force that one of the experimental forms of citation may have could be weakened by some subjects' belief that they need not worry about

receiving that form of citation in the future, after the experiment is completed. Avoiding this potentially serious threat to the value of the experimental results may suffice to meet the burden of justification carried by the deception.

Consider another example: the halfway house program intended to reduce recidivism among ex-addict parolees. Assume that the success of the program's therapeutic method may depend on maintaining an atmosphere of trust and support among parolees and members of the program staff, but that the program will entail deprivations of liberty greater than those associated with the status quo of straight parole. Participants' knowledge of the experimental nature of their treatment or of their random assignment to the program might result in resentment that could undermine the therapeutic atmosphere and result in failure of the program. The experimenters might therefore wish to deceive the participants by creating the impression that halfway house treatment was a new standard for ex-addict parolees. This would help ensure that the experiment afforded an accurate estimate of how effective the program would be if established on a routine basis. Nevertheless, such deception must bear an extremely severe burden of justification. The harm of random disparity involving deprivation of liberty would itself carry a very heavy burden of justification. To conceal that harm from those directly affected would preclude their challenging their treatment or knowing the reasoning believed to justify that treatment. The parolees would be left ignorant of delicate ethical judgments by which others rationalized harmful manipulations of their liberty. Such deception could be justified, if at all, only if the benefit to be derived from the experiment were extremely important and could be achieved in no other way. It would be difficult to meet the burden of justification in this situation.

CHAPTER VI. THE PROCESS OF JUSTIFYING PROGRAM EXPERIMENTS

This chapter considers the process of examining possible justifications for proposed program experiments in light of the harms analyzed in Chapter V. Our approach is to appraise the strengths and weaknesses of the arguments in order to help decision makers with their difficult task of balancing competing considerations and deciding whether an experiment, in its proposed form, is justified in light of its expected benefits.

Part A summarizes the questions addressed so far; their answers are necessary antecedents to a decision about the justification for an experiment.

A. Checklist of Conditions Precedent

1. Is the Proposed Experiment Within the Scope of This Report?

This report addresses only arguments about the ethics of experiments conducted within the justice system that will guide future change in the administration of justice. If a proposed experiment will not directly influence some part of the operation of the justice system, those considering the experiment should refer to the general literature on the ethics of research involving human subjects.[23] Experiments involving only simulation of justice system functions, for example, are not within the scope of this report.

There are of course cases of minor experimental innovation in the justice system where the issues are so minimal and it is so clear that no risk of significant harm is involved that it would be quite unnecessary to subject the decision to further scrutiny. Genuine doubt as to whether a particular proposed experiment does or does not fall within the scope of this report will always suggest that further scrutiny of some degree is required.

23. *See, e.g.*, The Belmont Report, *supra* note 3 and accompanying text.

2. Do Circumstances Warrant Considering Program Experimentation?

The status quo must need improvement in some substantial way. If the status quo has proved satisfactory so far, program experimentation may present greater risk than the alternative of forgoing the experiment. When the benefits of an innovative program can only be marginal at best—as in a program to reduce recidivism for a class of nonviolent offenders among whom recidivism is rare—the benefits of the experiment itself can be no more than marginal.

The proposed innovation must appear likely to be an improvement over the status quo. But at the same time, there must be uncertainties about the value or effectiveness of the proposed innovation that can be resolved with information from a program experiment. Experiments only provide information on measurable effects of an innovation; they may be unable to resolve other kinds of uncertainties about the wisdom of a proposed innovation.

Finally, experimentation must be the only practicable method of adequately resolving uncertainties about the effectiveness of the innovation. If strong rational grounds exist for predicting the effects of an innovative program without findings from a proposed program experiment, then there is a strong presumption against experimentation. The choice should then be between continuing the status quo and changing, without experimentation, from the status quo to the innovative program. Sometimes the needed information can be derived from an experiment that simulates the justice system's operation, or by analyzing information from similar programs or situations. If the needed information can reasonably be obtained in such a fashion, a program experiment probably cannot be justified.

These circumstances suggest several key ingredients of subsequent decisions. The nature of existing uncertainties about the effectiveness of the innovation will determine the nature as well as the precision of the information that potential experiments must be designed to produce. The extent to which the status quo needs improvement, as well as the proposed innovation's potential for making such improvement, will suggest how important it is that these uncertainties be resolved, and this will in turn define the potential strength of justification for harms associated with a proposed program experiment. (See Chapter II.)

3. What Experimental Designs Might Provide the Needed Information?

Consideration of alternative experimental designs will be guided by uncertainties about the relative effectiveness of the proposed innovation. Only those research methods that will produce information sufficient to remove these uncertainties need to be employed. It is often not necessary to employ the most rigorous experimental methods to resolve the relevant uncertainties. But methods that are simpler, less costly, and apparently less harmful to individual interests often will not provide sufficient information. In the broader view, simple methods may prove more costly and more ethically problematic than rigorous methods. (See Chapter III.)

4. What Ethical Difficulties Are Associated with Alternative Experimental Designs?

The range of experimental options must be considered in light of their scientific rigor, cost, inconvenience, and effects on the fundamental interests of individuals and society. This requires alertness to the burdens of justification carried by particular features of an experimental design. The nature and significance of disparity in treatment must be recognized, as must alternative designs that might involve less troublesome disparity, such as employing groups that do not share expectations of identical treatment. Because voluntary participation is always preferable to mandatory imposition, ways must be sought to maximize the voluntariness of both participation and nonparticipation. Any risks of infringing privacy must be minimized, and any plan to conceal from subjects the nature of their treatment and participation in the experiment must be considered with utmost care.

The process outlined above will allow identification of one or more plausible experimental designs, each of which will present a particular combination of potential benefits and harms. The benefits may be of several types: information that is useful in resolving decision makers' uncertainties, direct benefit to subjects from participation in the experiment, improvements in the justice system that may ultimately benefit those subjects, and improvements that may benefit some larger group from which experimental subjects are drawn. The harms may include infringement of individual interests and the risk that misleading experimental results will lead to unfortunate decisions regarding the innovation: it may be adopted when in fact it is ineffective, or it may be discarded when in fact it is superior to the status quo. (See Chapter V.)

With this range of considerations in mind, the final step is to evaluate, on ethical grounds, the asserted justifications for the experiments.

B. Decisions About Ethical Justification for a Program Experiment

Justifying a program experiment on ethical grounds requires evaluating the harms and benefits of available experimental designs, and on that basis deciding whether to experiment, to retain the status quo, or to adopt the innovation without experiment despite uncertainties regarding effectiveness.[24] Therefore the harms and benefits associated with each alternative must be weighed against those associated with other alternatives, both experimental and nonexperimental.

In making these judgments, one must be mindful of uncertainties associated with many of the possible harms and benefits. Although some experimental practices are certain to produce harm, the consequences of other practices are less certain—they involve risks or potential benefits. The use of sensitive information, for example, may not by itself entail harm, but it presents some risk that harm could result if confidentiality is breached. If a harm is not certain to result, but merely risked, the burden of justification should be discounted accordingly. Although some harms may be so severe that a bare possibility of their realization cannot be tolerated, in most cases the burden of justification is eased as the likelihood of the harm's occurrence is reduced.

Similarly, the potential benefits of an experiment or of the program experimented with must be discounted in light of the possibility that they will not be realized. Predictions about beneficial consequences of an experiment are subject to uncertainties surrounding the accuracy of the results, the influence of the findings on policy makers, and the effectiveness of the program in routine as opposed to experimental application. When weighing potential benefits as justifications for a particular experiment, these uncertainties must be taken into account.

The remainder of this part illustrates a range of important features of experiments to demonstrate the committee's view of how arguments about justification ought to proceed. The final part of this chapter considers outer limitations on experimental practices,

24. Full justification for undertaking an experiment depends on both the ethical propriety of the experiment itself and the authority of the administrator who would undertake the experiment. Here the focus is on ethical propriety; authority is discussed in Chapter VII.

recognizing that some practices may be impossible to justify in the context of program experimentation.

1. Where the Harms at Stake Are Modest

Consider a program devised to address the severe problem of recidivism among parolees who are chronic narcotics users. The initial requirement that the status quo must need improvement is easily satisfied. Assume that recidivism among such offenders is frequent and that there is good reason to believe that it results from the offenders' entrapment in a vicious circle. On being paroled, the offender encounters great difficulty in finding a job and any sense of security in the community. The resulting stress leads to a return to narcotics use and addiction, which in turn leads to theft to pay for narcotics. The offender is soon back in prison.

The program thought promising for improving this situation is a simple one. A number of concerned citizens have offered to serve as counselors to ex-addict parolees, helping them to find jobs, friends, a place to live, and generally a life without narcotics or crime. Each parolee would be assigned a counselor who would be available to help at the parolee's request. The program itself would be a service available to the parolee; it would involve no predictable harm. Moreover, because the program would be staffed by volunteers, it would not require public funds. The only uncertainty about the consequences of the program—a very serious uncertainty—is whether it would be effective in reducing recidivism among ex-addict parolees.

Now consider the merits of ascertaining this program's effectiveness through a program experiment. Regardless of whether the program were found to be effective or ineffective, there would be obvious benefit in having that information. If dramatically effective, the program surely should be expanded, and corrections officials would have discovered an approach to reducing recidivism that might be extended to other types of offenders. If the program were found ineffective, the experimental results would save the volunteers from inconveniencing themselves to no avail, and those concerned with reducing this kind of recidivism would benefit by learning that they must turn to other approaches.

Assume that there are two methodologically plausible ways of testing the effectiveness of this program: a randomized experiment and a before-after experiment. In the randomized experiment, parolees would be assigned randomly to two groups; those in one group would have counselors, those in the other group would not. Subsequent recidivism rates for the two groups would quite reliably

and precisely reveal how effective the program was. In the before-after experiment, all parolees would be assigned a counselor, and their recidivism would be compared with that of past ex-addict offenders. This comparison, for reasons given in Chapter III, could be relied upon only if it indicated that the program produced dramatic effects. The randomized design would entail harm in the form of disparate denial of benefit to the control group, while the before-after design would entail only the more modest harm of treating the parolees as objects of experimental study.

The choice between these two experimental designs will be guided primarily by the importance of obtaining the most reliable information in the circumstances. Because the randomized experiment would entail significantly more harm than the before-after study, it can be approved only if the before-after experiment would not provide sufficient information to guide future decisions. If those who were to administer the program would be warranted in continuing it even in the absence of strong evidence that it was effective, then the randomized experiment probably could not be justified. The before-after design would be sufficient as a prudent, though imperfect, check against the modest risk that the program will be counterproductive.

What about the alternatives of forgoing the program or adopting it without any prior experiment? Forgoing the program might be ruled out easily: the only harm it would avoid is inconvenience to those who serve as counselors, a "harm" that is voluntarily and perhaps gladly accepted. Adopting the program on a general basis without any prior study should also be unacceptable, because the before-after experiment offers the possible benefit of important and somewhat reliable knowledge but involves only trivial harm to subjects. The decision, then, will be between a before-after and a randomized experiment. Because the experimental program involves disparate allocation of benefits rather than harms, the burden of justification for a randomized experiment is mitigated. But if the need for maximum possible information about the value of the experimental program is relatively modest, the before-after experiment is probably the ethically superior choice. In cases, however, where the need for maximum possible information about the value of the experimental program is substantial, the case for choosing a randomized experiment in the face of such modest harms is, of course, much stronger.

2. Where the Harms at Stake Are Substantial

Consider a second hypothetical program addressed to reducing recidivism among ex-addicts: a halfway house program. This program would require that the parolee reside at the halfway house for the first six months of parole, observing its rules and participating in group therapy programs with other house residents. The parolee would be compelled to participate in this program, and the requirements of the program would be made conditions of parole, so that failure to cooperate might result in revocation of parole. The program would be established with public funds and would be quite costly.

Now consider the alternative ways of experimenting with this program. A before-after approach cannot be employed because not all parolees would be selected for the program (*e.g.*, some would be deemed unable to participate effectively in the therapy programs), and it would be impossible to ascertain from existing records who among past parolees would have been selected for it. (This is very often the case, as was noted in Chapter III.) A comparison between those selected for the program and those not selected would also be unsatisfactory, because the two groups would probably exhibit differing rates of recidivism in any case. An effective experiment will therefore require that two comparable groups be established, some qualified parolees being assigned to the program and others being assigned to the status quo of straight parole.

There are two ways of establishing the groups. One would be to establish a halfway house to serve only parolees in a particular geographic area and to use parolees residing elsewhere for a comparison group. All parolees would be screened for eligibility to ensure that both the experimental and comparison groups include only qualified parolees. The alternative would be a randomized experiment, similar to the comparison group experiment except that the group of all qualified parolees would be divided randomly, rather than by geographic residence, into experimental (halfway house) and control (straight parole) groups.

Notice first that the harms associated with these two methods are analogous in several ways. Both would involve the harm of mandatory imposition of the experimental treatment, and both would involve substantial experimental disparity affecting a very significant interest—liberty. Depending upon what the relevant alternative to experimentation is—retention of the status quo or instituting the halfway house program without experiment—the disparity may be, respectively, a disparate imposition of harm or a

disparate provision of benefit. These factors determine the burdens of justification that apply to both alternatives, and they are substantial burdens.[25]

How then do the alternatives differ? In the comparison group design, the groups that would be treated disparately may not share expectations of identical treatment that would be offended by the disparity. Different communities commonly have different resources and programs, and parolees might not think it unfair or unusual that the treatment they are given differs from that given parolees elsewhere. The randomized design, in contrast, would entail markedly different treatment of parolees in the same community, and thus might markedly offend their sense of fairness. Random disparity in the treatment of individuals from a single group and community does not conform to any common expectation, and almost always offends an expectation of identical treatment.

Yet it is precisely because groups from different communities may differ in ways other than their exposure to the halfway house program that the methodological soundness of the comparison group approach is uncertain. It may not be known whether eligible parolees from the two groups would exhibit similar rates of recidivism even in the absence of the halfway house treatment, so it would be difficult to infer the effectiveness of the program from a comparison of subsequent recidivism rates. The problem is not one of knowing that the two groups would differ in marked ways, or of knowing that they will be exposed to different influences that could distort the results of the experiment. Rather, the problem is that one cannot be certain whether differences in recidivism rates are the result of the halfway house program or some other unknown factors that influence the groups differently.

The randomized experiment would offer a more credible comparison, by better ensuring that subsequent differences between groups are attributable to the halfway house treatment. It would present some danger, however, that the subjects' resentment of their random assignment to the halfway house would undermine the effectiveness of the program. Such resentment could produce findings showing the program to be ineffective, when in fact it would not be in the absence of the resentment over the experiment itself. (There is little danger, however, that such findings would suggest the program is effective when in fact it is not.)

The choice between these experimental alternatives will depend on the weight accorded to the greater harm of random as opposed

25. As is discussed more fully at page 59, *infra*, this example presents a possibility for assigning parolees to the halfway house on a voluntary rather than a compulsory basis, which would make either proposed experiment substantially easier to justify.

to geographic disparity, on one hand, and the weight accorded to the greater benefit of more reliable information, on the other.

Finally, consider the alternatives to either form of experiment. Establishing the program on a general, nonexperimental basis would avoid disparity but leave the decision makers uninformed about the program's effectiveness. It would be very difficult to argue cogently that disparate imposition of harm is itself so offensive that one must prefer to impose the harm uniformly and in continuing uncertainty whether that harm is justified by any actual benefit. Forgoing the program entirely cannot be accepted easily because the program is thought likely to be superior to the status quo despite its cost and harms. If the status quo is believed to be seriously inadequate and the halfway house program is believed to promise significant improvement, a decision to forgo experimentation and retain the status quo may well be ethically unacceptable.

When the harms associated with a proposed program are great and the effectiveness of the program uncertain, it will usually follow that both the risks associated with proceeding in ignorance and the harms associated with rigorous experimentation are great. But if an experiment can resolve important uncertainties, great benefit may accrue—both to the justice system and to actual or potential participants in the experimental program.[26] A harmful and ineffective innovation will be abolished, or a harmful but effective program will be vindicated. In either case, the goals of our system of justice will be advanced, and future parolees will benefit—either by being spared the harm of an ineffective program, or by being afforded the benefits of an effective one. In these circumstances—where substantial harms are at stake—the choice will almost always be between conducting an experiment that will clearly resolve important uncertainties or forgoing the program entirely.

The need for very clear resolution of uncertainties will not necessarily require use of a randomized experimental design, however. In the example just discussed, for instance, a before-after design might be entirely satisfactory if circumstances permitted its application. If the experimental program were intended for application to a class of parolees that could be identified equally well in past and present parolee populations, and if an adequate measure of outcome—subsequent convictions, for example—could be applied to both groups, a before-after experiment might produce very clear results. When a before-after design will be adequate to resolve uncer-

26. Benefits that flow directly to the experimental subjects are given special consideration at pages 59-61, *infra*.

tainties, it will usually be the obvious best choice among experimental as well as nonexperimental alternatives.

3. Situations in Which It Is Possible to Obtain Consent

Consent is often not available as an element of justification for an experiment within the justice system. But the principle of respect for persons emphasizes the special value placed on allowing individuals to form their own judgments of how to weigh harms they may suffer against benefits they may receive. Therefore, it should always be asked whether, among the alternative research methods available, one can be found that will affect only individuals who have freely and fully consented.

Consent does not release the justice system administrator or the researcher from the obligation to justify the balance of harms and benefits. But it can remove one harm—that of mandatoriness. This report has emphasized the harms, dangers, and difficulties associated with mandatory participation in or exclusion from experiments. It is worth noting, however, that experiments can be conducted that benefit our system of justice and create none of the more serious ethical problems discussed in this report. These involve only subjects who have fully, freely, and knowingly consented to participation.

Consider, for example, a test of presenting trials by videotape. Videotape has been used successfully to present the testimony of witnesses who are unable to attend trial. Some have suggested that, at least in certain civil cases, benefits might derive from presenting entire trials to judge or jury by videotape. Proponents might argue that advantages include counsel's ability to piece together testimony taken at convenient times; the judge's ability to perfect the presentation of evidence and arguments by erasing improper questions, answers, and arguments from the videotape, rather than simply admonishing jurors to disregard what has been heard; and the opportunity for counsel to rehearse and polish their arguments at leisure.

Yet few would suggest that trials be presented by videotape as a routine procedure, either mandatorily for selected cases or even voluntarily, without careful prior analysis and experiment. A court might undertake an exploratory test of this concept, however, by soliciting counsel to volunteer cases for videotape trial, on the condition that all counsel and parties must concur.

In that situation, the exploratory nature of the program, its use only with the consent of all interested persons, and its availability in all cases in which it is desired are all that is needed to justify

the experiment. All harms that the research might otherwise inflict are avoided by obtaining the consent of every party involved.

Consent could also play an extremely important, although not wholly justificatory, role in the halfway house experiment discussed at pages 55-57, *supra*. If the halfway house treatment could be offered as an alternative to the last months of imprisonment rather than imposed in lieu of straight parole, subjects' participation could be made substantially voluntary without sacrificing the usefulness of the experimental results. Potential subjects could be offered the chance to participate in the experiment, subject to random assignment. Consent to participate would not be fully voluntary because many subjects would be motivated to participate simply to avoid continued imprisonment. The participants' view that the halfway house is an improvement over the status quo would reduce the harmfulness of the experiment and ease the required burden of justification. However, some potential subjects might not consent to the halfway house treatment, even as an alternative to imprisonment. This type of experiment would be incapable of reliably assessing the program's effectiveness for the latter group of potential subjects.

4. Situations in Which the Experiment May Benefit the Same Individuals It May Harm

An especially strong source of justification for harms in an experiment is probable benefits for the individuals harmed. Such benefits can occur either as a consequence of participating in the experimental program or as a consequence of information derived from the experiment. These outcomes illustrate particularly well the importance of viewing program experiments as segments of the larger process of the administration of justice. Here, that process is one whose outcome will be more beneficial treatment of the participants, that is, treatment yielding more efficacious results with no more onerous means, or the same results with less onerous means.

Participation in an experimental program is often likely to result in some tangible benefit to the participants, and this may be true even when such participation has fundamentally harmful aspects. Mandatory assignment of parolees to a halfway house program that involves greater restrictions of liberty than would otherwise apply, for example, is harmful to their immediate interest in liberty. But if participation in the program promises to help the parolee avoid narcotics addiction or future recidivism, it may afford the individual substantial compensatory benefit. Although the mandatory nature of the parolee's participation requires that the program

be viewed as harmful, it does not require that potential benefits be ignored in the calculus of decision. Clearly, a harmful experimental treatment that is intended to benefit the persons it harms is preferable to a treatment that harms participants primarily for the benefit of some other or wider group. Indeed, where the experimental treatment will entail very substantial harm to participants, a high expectation of benefit to those harmed may well be essential to meet the requirement that probable benefits must clearly outweigh anticipated harm.

Similarly, imposing mandatory arbitration as a prerequisite to trial in civil cases must be regarded as harmful to the interests of litigants. But it is intended to reduce costs and delays and thereby benefit the litigants it is imposed on, and the likelihood of that benefit should be recognized in evaluating the ethical propriety of the proposed experiment.

Much less common are experiments whose results, as distinguished from participation in the experiment, may directly benefit persons harmed in the course of the experiment. Consider an experimental test of a special program for persons sentenced to long prison terms: the relatively costly construction of a prison unit designed to give inmates a substantial sense of personal privacy and community within the prison walls, and thereby reduce inmate violence. The hypothesis is that these inmates have little to lose from violent and disruptive behavior; if they were given private apartments and modest amenities, they might acquire a sense of identity and worth that would reduce their motivation for violence. Because the program will require costly physical remodeling of a prison unit, it is first to be tested on a limited, experimental basis. Inmates not chosen to receive the experimental treatment will be harmed by disparate denial of a substantial benefit.

The harm those inmates suffer may be justified more easily than in most situations. The costs of the special program cannot be accepted without reliable evidence that it will help reduce inmate violence. If the program is effective, its consequent adoption on a general basis will ensure that the same individuals who were denied the benefit for purposes of experimentation will thereafter receive it; without the experiment, no prisoner would receive it.[27]

The situation differs when individuals encounter the justice system for relatively brief periods. If they suffer harm occasioned by a program experiment, their encounter with the justice system

27. The possibility that general adoption of the program might be precluded by fiscal considerations would, of course, undermine the possibility that those harmed would eventually receive benefit. But a basic premise of this report is that it would be wrong even to consider the hypothetical experiment unless the program were seriously intended for general adoption. (See pages 11-12, *supra*.)

is likely to end before the benefits of the experiment are converted into general practice. Benefit purchased at the cost of harm to their interests will generally accrue only to other individuals at some future time.

Since it is uncertain whether experimental subjects will have recurring contact with the justice system, it is equally uncertain whether, and to what extent, they might benefit from the results of an experiment they participate in. An experiment whose subjects are private civil litigants is unlikely to produce subsequent benefits for the actual subjects, because most civil litigants have very infrequent contact with the courts. On the other hand, the recurring participation of attorneys in the justice system makes it likely that they will be in a position to benefit from any changes that ensue from an experiment.

5. Where the Status Quo Produces Harm Similar to the Experimental Harm: A Special Case

Sometimes, a program experiment may be employed not to test the effectiveness of a new program, but rather to test the relative effectiveness of two or more programs that are already in routine use. Significant uncertainty about which of several current programs is most effective, or which is most effective for particular types of cases, may present both reason and opportunity to experiment.

Consider, for example, a juvenile court that has used for some time two special programs for certain types of juveniles who commit offenses against property. One program is probation coupled with a requirement that the juvenile attend weekly psychotherapy sessions. The other is probation coupled with an obligation to make restitution—to be fulfilled by earnings from weekly community service work. Offenders are assigned to one of these programs when more conventional alternatives—straight probation or detention—are deemed inappropriate. These are juveniles for whom straight probation has failed to produce any improvement in behavior, but for whom detention is believed to be too severe in light of the petty nature of their repeated offenses.

Assume that the decision to assign an offender to one of these programs is always individualized—that is, it represents a judge's good-faith effort to select the sentence most likely to curb the juvenile's delinquent behavior. In choosing between the psychotherapy and restitution alternatives, the judges often feel quite certain that a particular offender will best be served by one or the other. In many other cases, however, the judges are uncertain about which

alternative will be most effective. Each judge resolves these doubtful cases according to some personal principle. Some resolve doubt in favor of psychotherapy, feeling that it is more positive in its helping approach, and hence less onerous; others favor restitution, feeling that it implies more respect for the juveniles by making them responsible for their actions. Still other judges resolve doubt by acceding to the individual offender's expressed preference.

Although there is evidence that each program has been very successful in some cases, there is no systematic knowledge about which program works best for which offenders. Moreover, recidivism is a serious and continuing problem among this class of offenders. The judges are concerned that the program to which they assign an individual is often ineffective for that individual, when the alternative program would have been effective. They wish to test the relative effectiveness of the two programs in order to determine whether one is generally more effective than the other or, if not, which program is more effective for particular types of offenders.

In this example, there is no identifiable innovation to be tested. Instead, there are numerous innovative ways that the existing psychotherapy and restitution programs could be used. For instance, one or the other could be abolished entirely, or each could be used according to standards different from those that have heretofore been employed. What the judges seek is information that might enable them to devise more effective standards for the use of existing resources. Experimentation might be regarded as a means to produce an innovation.

Assume that the judges have already tried to improve their understanding of these programs' effectiveness by comparing recidivism rates among offenders assigned to the two programs in past years. Although sophisticated techniques of statistical analysis were employed, the results of this research were inconclusive because the researchers could not determine what systematic differences existed between the two groups of juveniles assigned to the different programs. There was simply no way to determine whether differences in the groups' recidivism rates resulted from differences in the programs' effectiveness or differences in the characteristics of individuals assigned to the programs. The only way to obtain clear information is by conducting some type of experiment in which comparable groups of offenders are assigned to the two treatments.

The threshold conditions for considering experimentation are met by the apparent inadequacy of the status quo, the possibility of improvement through experimentation, and the lack of available nonexperimental means to produce improvement. The apparent inadequacy of the status quo arises from the judges' serious doubt about how best to use the two programs—not from a mere lack of strong or scientific evidence.

Assume that there are only two experimental designs adequate to produce the desired information. Both would require that offenders be assigned randomly to either psychotherapy or restitution. In the first design, the population randomly assigned would consist of all offenders for whom the judge believes one or the other program is appropriate, including those offenders for whom the judge believes one of the programs is clearly preferable. In the second design, the population for random assignment would include only those offenders about whom the judge is uncertain—those cases in which the judge would ordinarily reach a decision by resorting to a personal principle for resolving doubt.

What burdens of justification do these designs carry, and how might those burdens be met? Both designs entail substantial harm: each would require that the judges abandon attempts to individualize assignments, and each would entail random disparity at the individual level (as opposed to randomization of groups or courts, for example). The first procedure carries the greatest burden of justification, because it would ignore the individual differences that make some juveniles clearly seem to need one treatment rather than the other. The first procedure might promise to yield greater knowledge than the second, because it would test the effectiveness of the programs for all offenders who might be candidates for either program—not just for that subgroup about whom present choices are especially doubtful. But it is unlikely that this procedure could be justified. The greater harm of the first procedure would have to be justified by added benefit, namely, resolving substantial uncertainty. Yet there is relatively little uncertainty regarding those offenders included in the first design who clearly seem to need one program rather than the other.

One might argue that the second design is easily justified, because those offenders who would be assigned randomly are already being assigned in a manner that is essentially random. But the present assignment is by no means random. Although the results of the current assignment procedure might be indistinguishable from the results of a random process, the procedures differ in a crucial way. The attempt to assign treatments on an individualized basis has substantial ethical value, so the choice to randomize carries a great burden of justification even though the two procedures may have similar results.

The hypothetical experiment might meet that burden, however, if it promises to help improve the results of future individualized choices. The results of the experiment may produce important

benefits, not only to society in general but also to the individuals who serve as subjects. Because the problem the proposed experiment addresses is recidivism among a population of offenders that includes the subjects, it follows that the subjects are likely to benefit in the future from better-informed sentencing decisions.

But, particularly in regard to random assignment applied to decisions as crucial as sentencing, the decision to experiment requires extreme caution. The responsibility to make individualized judgments contains a presumption that such judgments will best serve accepted ideals. That presumption should not be overridden simply because it cannot be proved, nor should mere difficulty in exercising that responsibility become the reason for experimentation. And the possibility of experimentation should not motivate a premature conclusion that distinctions cannot adequately be made between relevant characteristics of subjects.

C. Outer Limitations on Experimental Practices

In evaluating the justification for proposed experiments, we have focused so far on weighing harms against benefits. But certain courses of action are prohibited by our commitment to the ethical principles that underlie our system of justice, because they involve a kind of harm that no benefit can outweigh.

Certain measures would be ethically and constitutionally impermissible if adopted on a general nonexperimental basis; for example, the use of torture, denial of the privilege of habeas corpus, punishment through attainder, or unreasonable searches and seizures. Practices that are prohibited generally apply equally as absolute constraints on experimentation.

Even when no such absolute constraints are encountered, there are outer limits on what may ethically be justified in the weighing of harms and benefits. The principles of equal treatment and respect for persons, whether these are regarded as entering into the assessment of harms and benefits or as independent checks on that assessment, will serve as safeguards against any narrowly conceived calculus of efficiency or effectiveness. A program experiment can be deemed unjustified even though all standards we have set forth are satisfied—even though the harms are in some sense clearly outweighed by benefits—because the harm is simply unacceptable in the context of program experimentation.

In considering when these outer limitations occur, it is useful to differentiate among the kinds of harms or risks that may arise in an experiment. These cover a range including disparate denial of benefits, inconvenience, slight deprivation, substantial suffering (as

in certain types of treatment for narcotics addiction), and permanent physical or psychological harm (as in certain types of treatment for pathological violence). In experiments that involve mandatory and disparate infliction of substantial suffering, justification must be based on corresponding benefits to the experimental subjects themselves, and not merely on benefits to a separate wider group. More modest harm may be justified by possible long-term benefits to a wider or different population, but even in those cases the benefits must be so substantial as clearly to outweigh the harms. Mandatory participation in experiments that involve significant risk of permanent physical or psychological harm can never be so justified.

There will surely be disagreements about the boundaries of permissible experimental practices. Even the boundaries suggested above raise obvious questions: what harms should be counted in the category of substantial suffering? But it is not the task or intention of this committee to specify exactly where the lines of constraint should be drawn. What is most important is to recognize that there are limits on the harms that may justifiably be inflicted upon nonconsenting subjects of experiments within the justice system, regardless of the benefits likely to be gained. On some occasions, therefore, the crucial issue about justification for an experiment will not be the balance of harm and benefits, but rather the acceptability of the harm without regard to benefits.

CHAPTER VII. AUTHORITY AND PROCEDURES FOR UNDERTAKING PROGRAM EXPERIMENTS

This chapter addresses two concerns about the authority and procedures for program experimentation within the justice system. The first concern is whether an administrator who has the authority to undertake a program also has authority to undertake the program experimentally. The second concern relates to the procedures that may be needed to foster and guide responsible experimentation in the justice system.

Where does responsibility reside for the decision to undertake a program experiment? By definition, a program experiment entails some alteration in the actual operation of the justice system.[28] The term "responsible administrator" is used to refer to the officer or body by whose authority that alteration is made. Responsibility for a program experiment—for ensuring that it is ethically justified, properly authorized, and satisfactorily carried out—ultimately lies with the responsible administrator. Researchers may promote, design, conduct, and analyze program experiments, and they too will be concerned with ethical analysis; but ultimate responsibility rests with those who have the power to undertake the experiment.

A. Limits on Authority to Experiment

An administrator who does not have authority to implement a program on a general, nonexperimental basis also lacks authority to implement the program experimentally.[29] The opposite proposition, that authority for general program implementation includes the authority to experiment, does not necessarily follow. An experiment, unlike general policy, may have consequences—disparate

28. See the definition at page 3, *supra*.

29. Exceptions may be found in explicit statutory provisions. 42 U.S.C. § 1315, for instance, permits the Secretary of Health and Human Services to waive compliance with certain requirements of the Social Security Act for the purposes of experimental, pilot, or demonstration projects. It would not, however, be within the Secretary's authority to grant such waiver on a general, nonexperimental basis.

treatment, for instance—more harmful than those normally entrusted to the administrator's judgment.

All administrators are expected to exercise discretion within the limits of their delegated powers. Each federal district court, for example, is empowered by rule 83 of the Federal Rules of Civil Procedure to make "rules governing its practice not inconsistent with [the federal] rules." Similarly, a parole board is empowered to establish procedures and rules that govern the process of making parole decisions. The exercise of authority is not limited in such cases solely by the perimeters of formally delegated powers. It must also be limited by a sensitivity to the consequences of actions taken within the sphere of delegated authority. The authority of district courts to make local rules does not necessarily include authority to resolve all issues associated with experimentation involving those rules. A particular experiment performed within the framework of court activity might entail such significant disparity that, although justifiable in light of the analysis proposed in preceding chapters, it might exceed the bounds of administrative discretion entrusted to the court. In such circumstances, it would seem prudent to refer the matter for decision by an official or body that possesses a broader perspective or mandate.

The nexus between the level of authority and the importance of the interests involved may indeed have constitutional significance. *Greene v. McElroy*[30] involved a Defense Department program for the revocation of security clearances. The program, which lacked the normal procedural features of an adjudicatory hearing, was held invalid, even though the Defense Department was authorized to conduct security clearance proceedings and even though the procedures themselves would not necessarily have been judged unconstitutional had they been established by Congress or the President. The Court stressed its "concern that traditional forms of fair procedure not be restricted by implication or without the most explicit action by the Nation's lawmakers, even in areas where it is possible that the Constitution presents no inhibition."[31]

Similarly, in *Hampton v. Mow Sun Wong*,[32] a Civil Service Commission rule disqualifying aliens from employment in the federal civil service was held to constitute a denial of due process of law. The Court did not controvert the validity of such a rule if adopted by Congress or the President. The cardinal factor requiring higher approval or more explicit authority was "the quality of the interest at stake."[33] Whether or not comparable constitutional questions

30. 360 U.S. 474 (1959).
31. 360 U.S. at 508.
32. 426 U.S. 88 (1976).
33. 426 U.S. at 115.

are raised in the context of experimentation, the considerations underlying these decisions suggest the wisdom of resorting to more responsible authority as the justification for an experiment becomes less evident.

What factors should be considered in deciding whether a particular program experiment is within the bounds of the administrator's authority? The risks of assuming too much authority have been suggested; what are the risks of assuming too little? Changes in justice system programs cannot be limited to innovations certain to achieve improvement over the status quo. Such a limitation would leave administrators virtually powerless to address difficult problems. Administrators must undertake changes that entail some gamble; a new program is instituted because it is thought likely to result in improvement, even though there is some risk that it will create problems worse than those it is intended to remedy. Hence experimentation is a necessary element of the administration of justice. Improvement requires that new approaches be tried, accepting some risk that they may fail. Administrators must be encouraged to experiment with programs that have a good chance of succeeding, even though the success of those programs is sufficiently uncertain that rigorous experimentation is necessary. From this perspective, then, the administrator has more than authority to experiment; the administrator has an obligation to experiment. Overregulation of this effort would surely chill important innovation.

Armed with the authority and faced with the obligation to test innovations, how far should the administrator proceed before seeking an opinion or approval from a higher level of authority? The problem may be considered in two parts, one concerned with the experimental method, the other with the program to be tested.

Consider first the issue of authority in relation to the experimental method. A local probation official might quite properly establish a special unit whose sole purpose is to assist probationers in securing employment, professional counseling, schooling, or medical services—in general, to serve as a resource to help probationers overcome their difficulties. As sound as this program might appear, there is a real risk that it would be ineffective in reducing recidivism. If that were the result, the program would be regarded as an unacceptable use of the probation office's limited resources, and should be abolished. To resolve the uncertainty regarding its effectiveness, the program might be evaluated through randomized selection of probationers to receive the program services. The advantages of this approach must be weighed against the harm of affording disparate treatment to similarly situated probationers. Even though the experiment is justified on the ethical grounds set forth in the preceding chapters, and even with authority to establish the

program, the probation official's authority might not properly encompass decisions to create that sort of disparity, and the official must be sensitive to that possibility.

This caution is necessary because disparate treatment of similar individuals lacks important internal and political safeguards inherent in uniform treatment. A program applied to only a portion of the relevant population is less likely to generate effective political resistance than one that applies to all. Even when the generous motivation of a program that offers valuable assistance to all probationers is not open to serious doubt, affording that opportunity unequally might appear to be ill motivated—a form of favoritism disguised as experimentation. Any perception that it is improper may cast doubt on the integrity of the justice system. Because these consequences transcend the local program, it might be inappropriate for a local probation official to presume the authority to take such risks. Of course, this is not to suggest that the experiment should not be performed, but rather that it should be done with the express approval of a body or official with a mandate or perspective appropriate to the interests at stake.

The threats that experimentation sometimes poses for the integrity of the justice system also counsel against experimentation when the administrator's authority to undertake the program generally is in doubt. A program is not made less troublesome by being labeled an experiment; quite the opposite may be true. To undertake a program "experimentally" is to acknowledge uncertainty in regard to its value. If a program would push the limits of an administrator's authority even when its value was relatively certain, it must surely push them more when its value is unclear.

The limits on experimentation imposed by the level of an administrator's authority should not be interpreted as proscribing rigorous experiments while permitting either less rigorous but inadequate "experiments" or the general adoption of innovations of uncertain value. Rather, respect for these limits counsels that difficult decisions be brought to the attention of, and sanctioned by, those whose authority and perspective can lend assurance of integrity to the experiment, and can help to assure that the program under study is likely to be adopted if proven advantageous by the experiment.

Of course, there are numerous levels and structures of authority among the judges, legislatures, agencies, and officers who prescribe the programs, rules, and policies that constitute the administration of justice. Who possesses the requisite "higher level" of authority? This question is addressed in the next section.

B. Procedures for Undertaking Program Experiments

To this point, this report has focused on developing a conceptual framework for deciding whether to undertake an experiment within the justice system. In order best to ensure that experimentation facilitates improvement in the administration of justice, however, one needs more than a conceptual framework for decisions. Procedures and resources are also needed to foster responsible decisions both now and in the future. Appropriate procedures and resources will be determined by specific problems likely to be encountered.

Decisions made within the conceptual framework outlined in this report call for sensitivity to ways in which experiments may either infringe or promote certain recognized principles of our system of justice. The principles at stake are familiar to the administrators of the system. Administrators are less familiar with the application of these principles to the problems of scientific experimentation. Even less familiar is the logic of scientific experimental methodology, which in a fundamental way determines the ethical value of an experiment. A full appreciation of the issues that bear on decisions regarding proposed experiments requires administrators to have access to advice from persons skilled in experimental design.

With nothing other than a conceptual framework, judgments about the propriety of proposed experiments will be made on a largely *ad hoc* basis. Without access to the accumulated wisdom of prior judgments, making difficult decisions will never become less difficult, and a more comprehensive set of principles for experimentation in the law cannot develop. Those who must make these decisions need access to an evolving body of prior judgments.

Decisions to undertake experiments within the justice system that involve significant ethical or legal issues of the kind discussed in this report should be documented. Particularly within our system of justice, part of the justification for any action derives from public acknowledgment of both the principles at stake and the reasons that support the action.

These factors, as well as the need for approval of experiments involving harms that exceed the bounds of the responsible administrator's mandate, suggest the procedures and resources discussed below.

1. Advice

Administrators who must make decisions about the justification for a program experiment would benefit greatly from the advice of

persons experienced with both the methodological and ethical elements of experimental design. Some advice of this type is already available to particular components of the justice system. The Federal Judicial Center advises the federal courts, and the National Center for State Courts assists state courts and other state agencies. But research experts will not always be sufficiently sensitive to the needs of particular institutions.

Most justice system institutions would benefit from establishing standing advisory committees to assist individual administrators contemplating program experiments. Committee members, in addition to those with methodological expertise, should include diverse participants in the institution's operation—corrections and probation administrators, judges, attorneys, and perhaps members of the "subject" population. Such committees would develop and maintain institutional expertise about methodological and ethical elements of experimental design; they could serve effectively both to promote needed experimentation and to guard against unwarranted experimentation.

2. Approval

In many institutions of the justice system, no clear source may now exist for the kind of authority needed to approve ethically sensitive but important experiments. The need is for decision by a decision maker whose approval will help ensure that the judgment accommodates the varying interests of the institution, the experimental subjects, and the general public. In institutions with a clear hierarchical structure, led by an officer with substantial public mandate and accountability, decisions about experiments can properly be made as directed by that officer. For example, the Attorney General has a mandate that is probably sufficient to encompass almost any experiment that might be proposed within the Department of Justice. In federal and many state courts, by contrast, there is no equally clear line of administrative authority extending to an officer or body that possesses a similarly broad public mandate.

Suppose that a court wishes to experiment with an innovative local rule of court; it may not be clear that it has a mandate sufficient to warrant acceptance of harms that the proposed experiment would entail. In these circumstances, where should the court turn for proper approval of the experiment? The recognition of diverse interests that is needed for a satisfactory decision could normally be entrusted to a body composed of persons representing diverse perspectives on the institution: judges, attorneys with diverse types

of practice, social scientists, and perhaps members of the public. Bodies of this sort might be viewed as repositories of the institution's mandate for decisions about program experiments.

This approach is encouraged for all justice system institutions, including those such as the Department of Justice in which the locus of sufficient authority is already apparent. Ongoing oversight of program experimentation within an institution by persons who have differing interests in the experiments promises to provide valuable continuity and consensus about the need and proper methods for experimentation within that institution. Oversight responsibility might be lodged in existing bodies or in bodies established at whatever institutional levels appear most practicable.[34]

Although this committee encourages establishing such bodies for all justice system institutions, questions occur about the powers and functions of a reviewing or approving body. The "institutional review board" model,[35] which involves obligatory prior review of proposed experiments and veto power vested in the reviewing body, has been widely adopted in biomedical and behavioral research. But there is controversy about the value and effectiveness of that model.

Application of that model to experimentation in the administration of justice deserves particular scrutiny. On the one hand, were the review procedure to prove unduly cumbersome it might discourage valuable experimentation. This report has suggested that those in positions of authority in the justice system ought to try to improve the administration of justice. But that laudable goal will probably not be sought through scientifically designed experiments if administrators feel too burdened by a review process that does not pay due regard to the relative independence that many of them (particularly judges) traditionally exercise. If the process of review not only is mandatory in every case but carries with it a veto power, administrators who begin with no commitment to experimentation may decide to forgo innovation or to adopt new programs without prior experimentation in order to avoid the review process.

On the other hand, similar fears have turned out to be unjustified in the area of biomedical experimentation. Biomedical researchers today largely accept the need for prior review of the design and justification of experiments involving human beings. The failure of some physicians to proceed with scientifically designed experiments reflects their discomfort with the role of re-

34. We do not attempt to determine the appropriate bodies to exercise this oversight responsibility. Nor do we attempt to determine the extent to which existing institutions may already possess the necessary authority.

35. See 45 C.F.R. Part 46 (1981); 46 Fed. Reg. 8366-8392 (Jan. 26, 1981).

searcher vis-a-vis patients or the actual difficulty of conducting experiments much more than the burdens of obtaining institutional review and approval.

Physicians and others in the health field begin, however, with a stronger orientation toward scientific research than do justice system administrators. Thus, the body that offers advice and assistance in designing experiments will have the tasks not only of demonstrating the value of experimentation where needed and appropriate but also of facilitating the process of approval and review.

In some instances it may be appropriate for a single body—if broadly representative—both to foster research and to review proposed studies. Yet the two tasks are quite distinct. The ethical and methodological goals in designing a proposed experiment are to produce the least harmful but most effective design that is adequate to provide the needed information. But the ultimate decision about justification requires that a proposal be evaluated in a broader context: whether the benefit likely to flow from the experiment is sufficiently important clearly to outweigh attendant harms or risks. Persons who have invested effort in designing an optimal experiment may have difficulty accepting that the proposal could nevertheless fall short of justification.

The decision whether to have one body or two will thus have to be guided by local administrative circumstances and by whether the experiments pose issues of particular sensitivity that merit review by an entirely independent group. In the absence of a history of impropriety or actual abuses, the committee does not recommend that a separate review body (with power to approve, modify, or disapprove) be obligatory in every instance. It is our expectation, however, that when a proposed experiment entails especially weighty elements of harm and benefit, the administrator-sponsor and the body providing design advice will refer the experiment to another forum for final decision. In institutions where experimentation frequently involves significant harm or risk to individual subjects, it may be important to ensure generally and very cautiously against overzealous experimentation by vesting final authority for decisions about justification in an officer or group that has no special interest either in the design of program experiments or in the innovations they are intended to test.

3. Documentation and Publication

A decision to undertake an experiment involving significant ethical or legal issues should be documented by the responsible administrator in accordance with the conceptual framework recommend-

ed in this report. Such documentation should serve two purposes. First, the document will serve to clarify both the purpose of the experiment and the ethical considerations that have entered into the decision. Second, documentation will aid analysis of similar experiments proposed in the future. Decisions not to experiment often involve considerations that will be useful to those faced with similar decisions. If it can be done without imposing undue burdens on administrators, sharing of such information, including the reasons for the decision not to experiment, should be encouraged.

The documentation should include:

1. a statement of the legal basis of authority to undertake the program itself;
2. a statement of the circumstances that warrant experimentation, that is, explanations of the need to improve the status quo and of the uncertainties that the experiment is intended to resolve regarding effectiveness of the program;
3. a statement of the experimental designs likely to be effective in resolving those uncertainties;
4. an analysis of the ethical difficulties and justifications associated with alternative experimental designs;
5. a statement of the arguments according to which the chosen experimental design is judged to be the ethically superior alternative; and
6. a description of the process of advice and review to which the proposed experiment has been subjected, and, if the experiment is to be undertaken solely on the authority of the administrator, a statement why the administrator's mandate is sufficient to permit the experiment.

This report's necessary limitation to general guidance emphasizes the need for much more specific and detailed analysis that can only occur when actual experiments are proposed and analyzed. Documents produced as recommended above should be collected in a way that will permit reference to them as a body of informal precedent in the field of program experimentation.

The committee has no specific recommendation regarding how these decisions should be made available for reference as precedent. Perhaps a joint publication by the various institutions that foster and conduct experiments within the justice system would be valuable and effective. Establishing a suitable publication or repository for these decisions might involve such diverse institutions as the Federal Judicial Center, the National Center for State Courts,

the Department of Justice, the National Science Foundation, and the various state and federal courts, attorneys general, and corrections administrators. We urge initiative and cooperation among these parties in the task of developing a mechanism to collect and disseminate decisions about experimentation within the justice system.

CONCLUSION

Continuing sensitivity to the ethical problems of rigorous experimental designs must be balanced by sensitivity to the ethical problems of experimenting without effective design or of innovating without benefit of responsible prior experimentation. Our system of justice places great value on treatment of individuals in accordance with the principles of equal treatment and respect for persons; it also places great value on rational development of policy, which in turn can be realized through well-designed program experiments. These are not incompatible values. Responsible accommodation among them is necessary to improving our system of justice.

Scientific methods offer great promise of improving the administration of justice. Decisions to experiment involve complex and sometimes unfamiliar ethical issues and impose burdens, often severe, on the administrators who bear the final responsibility to make them. The approach suggested in this report seeks to make that burden manageable. By using institutional advisors, publishing decisions, and encouraging debate, the procedures we suggest will focus attention on these ethical issues as they arise on a case-by-case basis. Gradually the difficult issues may be resolved with more precision than is possible within the broad outlines presented in this report.

APPENDIX A

Text of the Chief Justice's Letter of January 24, 1978 to Committee Chairman Edward D. Re

Dear Ed:

I am writing formally to invite you to accept appointment as Chairman of the Federal Judicial Center Advisory Committee on Experimentation in the Law. As you know, I am also inviting other distinguished judges, scholars, lawyers, and representatives of public interest to serve with you.

The mission of the Committee will be to try to identify, define, analyze, and recommend resolution of issues bearing on the propriety, value and effectiveness of controlled experimentation for evaluating innovations in the justice system. Controlled experiments involve the random provision of disparate treatments. It is the most potent methodology for evaluative research—standard in medicine, education and psychology. We need to apply this concept to our problems even at the risk that its use in courts and other justice agencies may possibly raise constitutional and political questions peculiar to justice institutions. It is these questions with which the committee must deal. The ultimate purpose will be to provide guidance to researchers, judges and administrators who must decide what areas are appropriate for controlled experimentation.

The Center will provide supporting services within its resources. It is likely that prior to its final report, proposed recommendations of the committee will be aired before a conference of judges, lawyers, litigants and researchers—those for whom the report will have the most direct impact.

I am pleased you have accepted this assignment. You will find it challenging and rewarding.

Cordially,

Warren Burger

Honorable Edward D. Re
United States Customs Court
One Federal Plaza
New York, New York 10007

CONTINUED

1 OF 2

APPENDIX B

Methods for Empirical Evaluation of Innovations in the Justice System

Prepared for the Committee by E. Allan Lind, John E. Shapard, and Joe Shelby Cecil

Committee Note

This appendix was originally prepared to afford committee members an understanding of the theory and basic techniques of experimental research design. The committee recommends it to justice system administrators as it was offered to us—as a means to enhance their understanding of the capabilities, limitations, and logic of experimental methods, and thus to enable them to work effectively with researchers they may call upon to help design and execute a program experiment. We should emphasize that the administrator can and should play a central role in decisions concerning the design and execution of a program experiment. The research expert may be indispensable in the effort to *foresee* potential problems that may produce ambiguous experimental results, and to devise methods to circumvent those problems. But the administrator is responsible for *judging* the potential consequences of such problems for future policy decisions. This calls for a basic comprehension of the theory of research design, which we believe is readily available from this appendix.

It is not within this committee's competence or mission, however, to endorse the scientific correctness of the schools of thought underlying the material presented here. There are controversies among research methodologists that touch upon the relative importance of certain of the matters the appendix addresses. The appendix reflects substantial contributions from major established schools of thought. We recommend the appendix as an introduction to the concepts of methodology, and believe it will aid the justice system administrator in making sound judgments about the advice and recommendations of research experts, regardless of those experts' particular school of thought.

TABLE OF CONTENTS

I. INTRODUCTION

This appendix provides an overview of empirical research methods used to assess the effects of an innovation in the justice system. Aside from techniques for preserving anonymity or confidentiality, all material in this appendix addresses the discovery of cause-effect relationships. The discovery of a cause-effect relationship between the innovation and the characteristics it is to affect is the principal goal of empirical evaluations. Section II discusses the construction and timing of interventions and observations to increase the likelihood that a study will yield unambiguous information on whether an innovation *caused* a particular effect; this is typically referred to as the "research design." Section III concerns measurement of potential program effects; it deals with such issues as sources of error in measurement and the crucial question of whether an evaluation can speak at all to the questions on which ultimate policy decisions must be based. Section IV discusses interpretation of the results of an evaluation, focusing on issues that must be considered to give meaning to the raw data of the research. Section V discusses means to preserve the privacy of individuals studied in the course of the evaluation and the confidentiality of their responses and comments.

Before presenting the substance of this appendix, it is in order to issue both a reassurance and a caveat. This appendix is intended to be fully comprehensible to readers with no special expertise in research methods or in statistics. All of the points made here derive from the application of sound logic to the consideration of potential problems in evaluation research. The issues addressed are not simple, but neither are they so esoteric as to be beyond the understanding of the diligent, but uninitiated reader. We hope that this document will convey the knowledge necessary for a justice system administrator to consider and act intelligently on questions of ethics and to collaborate effectively with methodologists to assure a reasonable and informative evaluation.

However, it should be stressed that this appendix is an introduction to, rather than a complete treatment of, research methods. The basic concepts underlying empirical research are presented here, but the finer points are necessarily beyond the scope of a document such as this. Good evaluation depends on a close collabora-

tion between policy makers and evaluation specialists, and expert assistance is essential to apply the concepts presented below.

A Hypothetical Example

In order to give some additional continuity to the following presentation and to relate abstract issues in evaluation methodology to concrete questions of the sort that arise in any empirical study of policy change, we will often use examples based on a single hypothetical program. For the purpose of these examples, suppose that an administrator in the federal prison system wants to test a program of special, intensive training in job skills for inmates who are about to be paroled. Suppose further that the program, *if successful*, is expected to increase the prospects of regular employment among parolees and to result in lower rates of recidivism. And suppose that the program is expensive and the administrator has decided that clear evidence of its effectiveness is necessary before such training will be made a permanent and widespread feature of federal prisons.

We stress that these examples are purely hypothetical. In using them for the purpose of illustrating methodological concerns, we posit various fact situations and various actions on the part of the administrator and the program participants. But we do not intend to convey that such fact situations would arise in a real program of this sort or that the actions we suppose are proper for an evaluation.

II. RESEARCH DESIGN

The design of an evaluation is the overall strategy for extracting information from the test of a program. Although a great deal of scientific technique and experience has been developed in this matter, it is important to recognize that there is *only one*, quite simple matter that is the goal of all experimental design: to assure that the comparison upon which an inference of causation may be founded is in fact a sound comparison. Notice that the concept of a cause-effect relationship does imply a comparison. To assert that the hypothetical job skills program causes increased likelihood of regular employment and decreased risk of recidivism is to say that participants in the program behave other than they would have without the program. This assertion is equivalent to saying that, all else being equal, a parolee afforded the program is more likely to be employed and less likely to recidivate than a parolee not afforded the program. The object of research design is to construct a study that approaches, as closely as ethics, practicality, and ingenuity allow, the "all else being equal" specification needed to infer causality.

The necessary comparison for inferences of causality can almost never be achieved by affording one individual the program under study and withholding the program from another individual. This is so for two reasons. First, when one is dealing with phenomena as complex as those involved in the success or failure of a social program, the variation that always exists between two individuals is so great that one can never be sure that differences in subsequent outcomes are not due to individual differences rather than participation in the program. To find that one parolee, who was given the hypothetical job skills program, found regular employment, while another, who was not given the program, did not is not conclusive because we know that there are many factors other than the program that might affect the likelihood of employment. Second, in nearly all justice system programs there is no expectation that an outcome will always occur when the program is afforded and will never occur when it is not. In the hypothetical program, the administrator hopes that the training will increase the likelihood of regular employment, not that it will render this potential benefit a certainty for every individual given the training.

For these reasons, evaluation studies must usually consider the outcomes of a program for groups of individuals exposed to the program in comparison to the outcomes for groups not exposed to the program. By studying data from groups, it is possible to generate summary statistics that give information on the general effects of the program and that allow statements to be made about overall consequences that might be expected if the program were generally available and routine. Thus, the essence of the research design becomes the construction or identification of groups that, upon observation, will yield the information needed to determine the effects of the program. In the hypothetical job skills program, the task is to arrange a comparison between a group of soon-to-be-paroled individuals who are given the training and a similar group who are not given the training in such a way that the clearest possible picture of the consequences emerges.

The quality of any research design lies in its capacity to eliminate or reduce the possibility of any explanation of the outcomes observed other than that the program caused the outcomes. Thus, any design can be assessed by adopting a skeptical frame of mind and seeking credible *rival hypotheses* that would explain potential results of the study without supposing that the results are due to the program. To the extent that credible alternative explanations

of potential results exist, the evaluation will be ambiguous and the design will be weak. For example, if rival hypotheses can be advanced for any apparent effect of the job skills program on employment or recidivism, the conclusion that the program caused the effect is rendered less credible. Although, as will be seen below, no design can eliminate all rival hypotheses, some designs are particularly good at producing relatively unambiguous results while others are likely to lead to ambiguity.

It should be noted that the strengths and weaknesses of research designs that assess cause-effect relationships are important only insofar as the value of an innovation is to be judged by observable effects and only insofar as these effects are sufficiently subtle to require careful comparisons if they are to be detected. The necessity for strong research designs depends on the questions posed by the policy decisions that must be made. It may be that the effects of a particular justice system innovation are either so clearly evident or so unclear in import that cause-and-effect evaluations are not necessary or relevant to the policy decision. Thus, for instance, abolishing diversity jurisdiction may have certain quite obvious effects; the crucial empirical questions might relate to public satisfaction with such a change. On the other hand, consider a change to six-member juries. If this change resulted in smaller recoveries by plaintiffs, we would have no sound basis to determine if that result is good or bad (unless the policy behind the change calls for no change in verdicts). Thus, an evaluation to assess the consequences of jury size for verdicts might have little to contribute to the ultimate policy decision. In such instances, careful description of the innovation and assessment of the perceptions of the public or of actors in the justice system, not inferences of cause and effect, might be the elements of proper evaluation. There are, however, many innovations for which policy decisions are contingent on information that can only be obtained by the use of strong research designs.

We turn now to the discussion of specific designs for evaluation studies.[36] First we present a class of designs that employ lotteries to construct groups that closely approach the "all else being equal" criterion of the crucial comparison from which causal inferences can be drawn. Methodologists call these designs "randomized experiments." We then present a second group of designs, termed

36. The material presented in the remainder of this section is based primarily on D. Campbell & J. Stanley, Experimental and Quasi-Experimental Designs for Research (1966) and T. Cook & D. Campbell, Quasi-Experimentation (1979). The bibliography at the conclusion of this appendix lists these and other standard works in the field of experimental design, which are recommended for readers who seek a more thorough and detailed discussion.

"quasi-experiments," which do not use lotteries, but instead rely on observation of preexisting groups or on observations before and after a program is instituted.[37] The designs presented below are not the only ones that exist; they are selected to convey the major options and issues that are often considered in determining which design is best to evaluate a justice system program.

Randomized Experimental Designs

Simple Randomized Experiment

Consider the following strategy for evaluating the hypothetical job skills program. Suppose we select 200 inmates from the group of potential participants in the program, twice as many as we intend to place in the program. A lottery randomly assigns half of these individuals to the program; the remaining half are given the normal, or status quo, treatment. Suppose further that we monitor, for a set period of time, both the 100 parolees assigned to the program and the 100 assigned the status quo treatment. Upon completion of this data collection period, we compare the percentage of program participants who hold regular employment and the percentage who have been convicted of a subsequent offense to the corresponding percentages in the group that did not participate in the program.

The random lottery[38] is the crucial element distinguishing this class of designs from all others, and the implications of its use in the study are great. The assignment of any particular person to one group or the other is the result of a purely random process, and not the result of any characteristic of the person. If the groups are sufficiently large, the laws of probability assure us that it is very unlikely there will be any substantial difference between the group exposed to the program and the group not exposed to the program. The "all else being equal" criterion has been achieved, at least at the time the lottery is conducted, and if the only difference in how the groups are subsequently treated is the job skills pro-

37. In this appendix, we use the word "experiment" in the names of designs as it is used by methodologists as a term of art, rather than in the more common fashion used in the body of the report. In all other contexts, the word is used as it is in the report.

38. Random assignment to groups should not be confused with random selection of all subjects from some larger population. Random assignment assures that the two groups being studied are equivalent; random selection assures that the groups will be representative of the larger population. We do not consider the use of random selection in this appendix. We make the distinction here only to avoid confusion of these two techniques.

gram, we can confidently attribute to the program any differences we later observe between the groups.

The essential characteristics of a randomized experiment can be diagrammed, using a system of notation we will employ in presenting all the research designs we discuss. (We present such diagrams for each design we discuss in order to restate in formal fashion the major characteristics of the design. The diagrams convey only what is presented in the text, however, and can be ignored by readers who find them confusing.) The diagram presents the events that make up the design; events diagrammed to the right of others are later in time. Methodologists use the term "experimental group" to refer to the group exposed to the program, and "control group" for the group not exposed to the program. The particular design diagrammed here is termed the "simple randomized experiment" to distinguish it from the more complex designs discussed below.

SIMPLE RANDOMIZED EXPERIMENT

	P	O
R	.....	
		O

R		indicates two groups constructed by random assignment
P	=	exposure of one group to the program
O	=	observation of each group

The groups constructed by random assignment must be sufficiently large to allow the laws of probability to function to eliminate differences. Just as one would be more confident in predicting that 100 flips of a coin would result in something close to 50 heads than in predicting that 10 flips would result in 5 heads, so too one would be more certain that random assignment of 100 individuals would be more likely to eliminate extraneous differences than random assignment of 10 individuals. The number of individuals needed for a study depends on the variability of the characteristics to be examined, and statistical "power" formulas can be used to estimate this number if those who design the program can specify the minimum effect to be detected. If, for example, the administrator in our hypothetical job skills program can state that the evaluation of the program should be able to detect changes in recidivism of 15 percent or more, an evaluator can determine how many individuals should be included in the experimental and control groups.

Even in the absence of such definite estimates about the magnitude of the potential effects, statistical conventions exist that can provide guidance concerning the number of subjects that should be included in the study. (There is always a chance that the effects of a program will be so small in comparison to natural variation in the characteristics of interest that no firm conclusion can be drawn, and thus that the evaluation will not speak to the existence of the effects; this is a problem that can occur in any design, randomized or not.)

Statistical procedures can be used to assure that the results of a randomized experiment are not due to some fluke in randomization. In all designs, randomized or not, statistical tests are used to determine whether an apparent difference is sufficiently large, in comparison to natural variation in the data being collected, to permit the inference that the difference is real and not the result of such variation. Tests for "statistical significance" allow one to place a stringent burden of proof on the conclusion that the experimental and control groups do in fact differ and that the observed effects, if they exist, are not the result of random variability.

In the simple randomized experiment described above, we would be able, if we found statistically significant differences in employment and recidivism rates favoring the experimental group, to conclude that the program had caused these differences. Testing the strength of the design by adopting a skeptical approach reveals few alternative explanations and shows the randomized experiment to be relatively unambiguous. The results could not be due to preexisting differences between the groups, because the lottery has eliminated such differences. Nor could the results be readily ascribed to such factors as the economic climate in which employment is sought, because both the experimental and control groups are subject to the same situation in that regard.

In the randomized experiment, sound conclusions hinge on various assumptions. We must assume that assignment to the two groups was indeed random. If those in charge of the assignment have deviated from the use of a truly random procedure, the logic of the randomized experiment cannot be applied. Deviations from random assignment might occur, for example, if those assigning inmates to the job skills program thought that they should occasionally determine who was most in need of the program, a practice that would subvert the intended assignment scheme. Even if an apparently arbitrary, but not truly random process, such as assigning every other inmate on a list to one group, is used, the experiment would be suspect because, for example, the assignment scheme might be detected and manipulated. If such deviations did occur, the experimental group and the control group could not be as-

sumed to be equivalent, because systematic differences other than exposure to the program might have been introduced.

All designs, randomized or not, suffer potential weaknesses aside from those that may arise from preexisting systematic differences in the groups compared. If there are differences other than the program in the postrandomization treatment or environment of the experimental and control groups, a randomized experimental design may lead to erroneous conclusions. If, for example, the prison officials charged with administering the experiment felt that the control group inmates were being deprived by not being afforded the job skills training and tried to "make it up to them" in other ways, the desired comparison between the program and the status quo treatment would not be produced. Similarly, if the control group inmates knew of the program and felt that they were being unjustly deprived of it, they might behave in a different fashion, and the comparison would be invalid. (We return to this issue in Section IV below, in our discussion of "reactivity" and its implications for evaluation.) Considerations such as these show that the randomized experiment is not an infallible technique for evaluation; there are certainly situations in which it can lead to erroneous conclusions. The crucial issue in deciding the methodological attractiveness of a research design, however, turns on whether it is less fallible than available alternatives.

Although we have been discussing random assignment of individual persons to experimental and control groups, it should be noted that other entities can be randomly assigned and many of the logical qualities of the randomized experiment retained. It is possible to randomly assign cases, courts, or institutions, if enough are available to afford some reasonable expectation that the lottery will eliminate preexisting differences. For example, if twenty federal prisons were available for a test of the job skills program, ten could be randomly assigned to an experimental group, with all eligible inmates receiving the program, and ten could be assigned to a control group, without the program. Although, strictly speaking, the statistical analyses would have to be conducted using prisons as the "subjects" of the analysis, few methodologists would quarrel with application of the findings to inferences about the effects of the program for individual parolees. The use of random assignment of entities larger than the individual person or case is preferable when the potential effects of the program involve characteristics of the larger entity. If an evaluation seeks to determine whether a new program leads to increased civil filings in district court, for example, a design that included random assignments of courts to the experimental and control groups would be preferable to a design that randomly assigned cases.

Multi-Group Randomized Experiment

More elaborate randomized experimental designs are available for testing finer issues than whether the presence of an entire program causes some effect. Suppose the hypothetical job skills program involved not only skills training, but also some extensive assistance in securing employment after parole. The administrator might wish to know whether both elements of the program were needed to produce the desired effect. This information could be generated with a "multi-group randomized experiment" that used a lottery to create not two but four groups. One of these groups could be assigned the full program of skills training and job search support, the second group could be assigned a modified program that provided only skills training, the third group could be given only job search support, and the fourth group could be given the normal, status quo treatment. This design would allow the evaluation to determine not only whether the full program is effective (by comparing the first and fourth groups), but also whether it is the training or the job search support, or both, that cause the observed effects (by comparing the first, second, and third groups). In the diagram, rows with the notations "P_1," "P_2," and "P_3" indicate the various experimental groups with their different versions of the program, and the last row, with only observation and no program, indicates the control group.

MULTI-GROUP RANDOMIZED EXPERIMENT

	P_1	O
R		
	P_2	O
R		
	P_3	O
R		
		O

R ..... indicates the four groups constructed by random assignment

P_1, P_2, P_3 = exposure of three groups to different versions of the program

O = observation of each group

Before-After Randomized Experiment

If there is considerable variation in the characteristics thought to be affected by the program, or if the characteristics tend naturally to change with the passage of time, another elaboration of the simple randomized experiment might be useful. The "O"s that appear before exposure of the experimental group to the program indicate observation of all individuals immediately following random assignment to the groups and permit the evaluation to assess the change in characteristics of both groups from the time of randomization to the time the program effects are expected to occur. Because the groups are randomly constituted, however, it is expected that the "before" observations made on the two groups will be very similar, and the main reason for using the present design, rather than the simple randomized experiment, is that it is often easier to statistically detect differences in change in behavior than differences in the absolute level of the behavior.

BEFORE-AFTER RANDOMIZED EXPERIMENT

	O_1	P O_2
R	.....	
	O_1	O_2

R		indicates the two groups constructed by random assignment
O_1	=	observation of each group immediately following random assignment
P	=	exposure of one group to the program
O_2	=	second observation of each group

Suppose it were known that some participants in the program were in possession of marketable skills while others were not. The use of a before-after randomized experiment with measurement of skills in both groups prior to the delivery of training would permit more sensitive measures of improvement to be gained because individual improvement, rather than overall group differences, could be measured. The design would not differ from the simple randomized experiment in its basic logic, but it would differ in the sensitivity of the statistical tests that would be used to assess the results of the program.

In the following material, we will often compare other designs to the standard of the randomized experimental design, especially to point out the ambiguity that often plagues designs that do not use randomly constituted comparisons. Emphasis on the randomized experimental design stems from two related considerations. First, the randomized experimental design is a research strategy of such logical power that it is, from a purely methodological point of view, the ideal design in many evaluation situations. It is seldom the case that other evaluation designs test cause-and-effect relationships better than the randomized experiment. Second, because of its methodological attractiveness and its requirement of randomly created disparity, the randomized experimental design often poses the most severe ethical questions. Thus the randomized experiment is of particular importance as a prototype case in the ethics of program evaluation. The resolution of the questions raised by the randomized experiment is, of course, the topic of the body of the report, not of this appendix. Statements here about the relative advantages of randomized experiments are addressed only to *methodological* advantages; they take no account of ethical consequences.

Quasi-Experimental Designs

Quasi-experimental designs are those that focus on some "comparison" group of subjects not exposed to the innovative treatment, or that employ observations before exposure to the treatment, in order to infer what would have happened in the absence of the innovation. The comparison group, however, is always in some way systematically different from the "treatment" group,[39] in the sense that there is *some* identifiable difference between the groups other than the fact that one group receives the treatment and the other does not. These designs may yield ambiguous results whenever this systematic difference suggests a credible alternative explanation for apparent effects of the innovation.

Before-After Design on Individuals

An apparently straightforward, but actually quite problematic, design for evaluating programs that offer potential for some change involves constituting a single group of program participants, measuring the characteristics of interest, exposing the par-

39. The groups are termed "treatment" and "comparison" groups, rather than "experimental" and "control" groups, in keeping with a convention that distinguishes between groups in quasi-experimental and randomized experimental designs. The distinction has no other significance.

ticipants to the program, and then measuring the characteristics again.

BEFORE-AFTER DESIGN ON INDIVIDUALS

O_1	P	O_2

O_1	=	observation prior to exposure to the program
P	=	exposure to the program
O_2	=	observation following exposure to the program

In our hypothetical example involving the job skills program, one might use this design by recording for each participant whether he or she was employed during some period prior to the current conviction, and by observing the participant for an equal period of time after completion of the program and granting of parole. From this example, though, some of the problems associated with the design become apparent, if one adopts the skeptical approach required to test the strength of a research design. Suppose that it is found that parolees are indeed more likely to be employed following the job skills program than they were prior to their last conviction, and consider the alternative explanations that might be advanced. It is, of course, possible that the program has achieved its goal. But it is also possible that the greater likelihood of employment is simply due to the fact that the participants are older than they were at the time they were convicted and that older workers are more likely to find employment. It also might be the case that simply having been incarcerated has motivated the participants to seek employment. Or it might be that the increase is due simply to the normal monitoring of parolees, which might encourage them to find work. Another rival hypothesis, which would be especially credible if economic conditions had improved during the time of imprisonment, is that the job market has so improved that, with or without skills training, any parolee is now more likely to find employment. The capacity of the design to test the effects of the program is weakened to the extent that such alternative explanations are credible.

In general, before-after designs on individuals are subject to ambiguity whenever it is conceivable that changes would have occurred naturally in the characteristics of the individual, as the result of aging for example, or whenever it is conceivable that the results of the study might be due to changes in external circumstances between the time of the "before" observations and the time

of the "after" observations. Another potential source of ambiguity in this type of design arises when the observations that are repeated use a test that can be learned. For example, if the job skills program evaluation used a test of job skills before and after the training to assess whether the participants were actually learning skills, it is possible that simply knowing the nature of the test at the time of the "after" observation would lead to higher scores, even without learning new skills.

Finally, ambiguity can arise in this design when the individuals given the program are selected on the basis of their extreme position on some characteristic. If, for example, only inmates who have never had regular employment are eligible for the job skills program, one might see some improvement in their rate of employment simply because there may be a few participants who already had some skills but who have, by chance, never found work and because the others can certainly have no worse prospects than they did before the program. This state of affairs would lead to an apparent beneficial effect for the program, but this would be illusory because the change would have been evident even if no skills training had been given.

Thus, the before-after design may be useful in situations where it is judged that time-based changes are unlikely and where there is little possibility that erroneous conclusions can arise from the repeated measurement or from the selection of extreme groups. It is very often the case, however, that at least some of these sources of ambiguity are credible and pose potent threats to the conclusions of a before-after study. The before-after design on individuals is generally much less rigorous than a randomized experimental design. Although the randomized experiment has its own sources of ambiguity, these are generally regarded as less likely to pose serious threats. Of course, such overall comparison of the strength of designs has its exceptions, and the relative merits of any two designs must be weighed with reference to the particular program under study and with consideration of the credibility of the rival hypotheses that might arise in the application of each design to that program.

Simple Comparison Group Design

A second quasi-experimental design involves a comparison between two groups, as does the simple randomized experimental design, but uses groups that are known to differ in some systematic fashion other than exposure to the program. For example, the job skills program might be made available to eligible inmates at one

prison and the results compared to those seen with a similar group of inmates at a prison without the program.

SIMPLE COMPARISON GROUP DESIGN

	P	O
.....		
		O

.....		indicates two groups (*not* constructed by random assignment)
P	=	exposure of one group to the program
O	=	observation of each group

In the diagram, the broken line indicates the use of two groups, but the absence of the "R" signifies that the difference between the groups is not random. The group receiving the program is termed the "treatment group"; the group not receiving the program is termed the "comparison group."

Suppose that a group of soon-to-be-paroled inmates at Prison A is given the job skills training and a similar group at Prison B is not. Suppose further that it is subsequently found that the Prison A treatment group is more likely to be regularly employed and less likely to recidivate than the Prison B comparison group. These results would provide some basis for inferring that the program caused increased employment and decreased recidivism, but only if any potent alternative hypotheses could be dismissed. There is no assurance, for example, that inmates at Prison A do not normally find employment and avoid recidivism at a better rate than do those at Prison B. This might occur if the systematic difference that we know exists between the treatment and comparison groups is such that it affects employment and recidivism. It might be that differences in the nature of the inmate population, in the other programs provided at the two prisons, or in employment opportunities in the geographic areas to which the prisoners are released could explain the results of the study without reference to the job skills program.

This example points to the most troublesome aspect of comparison group designs: the possibility that differences existing before the program is instituted have caused or contributed to any differences that are subsequently observed. An evaluator may attempt to select a comparison group in such a way as to eliminate the most

obvious differences between the comparison and treatment groups, but there is always the possibility that remaining differences might provide a viable alternative explanation of the results. For example, if a variety of prisons are available for study, two might be selected that are similar in terms of inmate populations and programs other than the one under study, but there remains the possibility that some factor—for example, attitudes toward hiring parolees in the area to which most are released—could account for the results of the study.

A major practical problem that arises in many uses of comparison group designs is identifying the specific individuals to be included in the comparison group. Many justice system programs are targeted for certain groups, rather than applied across the board, and it may be difficult to set up a similar identification process absent the program. For example, if participation in the job skills program is contingent not only on incarceration at Prison A but also on the recommendation of a social worker that the individual would profit from the program, it may be difficult to know which of Prison B's inmates would have received such a recommendation had the program been in existence there. (Even if Prison B's social workers could be persuaded to replicate the recommendation process for the sake of the study, their selections, which would be known to have no consequences for the inmate, may not be similar to those made by Prison A's social workers, who would know that their selections might have substantial consequences.)

The comparison group design does eliminate one source of ambiguity that is a major danger in the before-after design. Because the comparison group is subject to the same general time-based changes as the treatment group, such changes are not viable rival hypotheses for any difference observed between the two groups. In our example of a before-after evaluation of the job skills program, we noted that an improvement in the overall economic climate might lead to the false impression that the program improved employment prospects for the participants. But in a comparison group design, such an error would be unlikely because both treatment and comparison groups are subject to the same overall economic situation and because we would require that the treatment group do better than the comparison group in order to conclude that the program was effective. It is important to note, however, that an entirely local change in economic situation could still lead to problems; if the economic climate improved in the area to which most of Prison A's inmates are released but not in that to which most of Prison B's inmates are released, it might appear that the program was effective when in fact it was not.

A common, but very problematic, use of the comparison group design involves comparing the treatment group to a group of individuals not selected for the program. In such evaluations, there is a troublesome contradiction between the assumptions made in the evaluation and the assumptions on which the program itself is based. The evaluator is assuming that the criterion used to select participants for the program does not affect the results, while the program designer presumably has included the criterion precisely because it is thought to affect the results. Consider the situation that arises if only inmates who have never held regular jobs are afforded the job skills program and the evaluation is based on comparison of their outcomes with those of the previously employed inmates not afforded the program. The program designer has included the criterion because it seemed reasonable that the group afforded the program would normally have more limited employment prospects, and the program might be effective even if it did not overcome all of this preexisting difference between the groups. If the program actually increases the employment prospects of the treatment group above what they would be without the program, but not enough to overcome the preexisting advantage enjoyed by the comparison group, the program can appear ineffective even though it is not. Conceptually similar problems can arise when one criterion for participation in the program is volunteering for it. Volunteers may be those who most need the program or those who expect to benefit from it. In either case, a systematic difference has been introduced that leads to a strong rival hypothesis.

In general, even the best simple comparison group designs are much less rigorous than the randomized experimental designs. The use of a comparison group design can be defended on methodological grounds only when one can be reasonably certain that the systematic difference between the treatment and control groups could not affect the outcomes being studied. Unfortunately, this is seldom the case.

Before-After Design on Institutions

One variation on the simple comparison group design is to compare the results observed with participants in the job skills program to the results that had been observed with similar parolees who were released prior to the test of the program. The vertical broken line in the following diagram indicates a group difference, but it is now a temporal difference, rather than a difference based on institutional or criterion distinctions.

As might be supposed by its resemblance to both the simple comparison group design and the before-after design on individuals, the

before-after design on institutions must contend with some of the problems of each of the quasi-experimental designs already discussed. Like the before-after design on individuals, there is danger that some general time-based change in the situation surrounding the program will produce an illusory effect (or an illusory absence of effect). Like the simple comparison group design, there may be substantial difficulty in identifying a truly similar "before" comparison group. Because it is often the simplest design available for the study of the effects of changes within a single institution, this design is probably the most common strategy for justice system evaluations. But the problems inherent in the design make it a strategy that risks inconclusive results.

BEFORE-AFTER DESIGN ON INSTITUTIONS

O_1	⋮	P	O_2

⋮ indicates two groups (separated in time; *not* constructed by random assignment)

O_1 = observation of the first group

P = exposure of the second group to the program

O_2 = observation of the second group

Because the before-after design on institutions can be used to attempt to provide a comparison on the basis of standard records about cases or individuals who were involved with the institution before the program began, it is tempting to employ it to evaluate a program. In such instances, the problems of identifying the proper "before" comparison are often so severe as to be insurmountable. Consider the problem of evaluating the job skills program using the before-after design on institutions, and assume that the program has moderately complex selection criteria for participation. The treatment group can, of course, be readily identified and its outcomes measured, but the attempt to identify which of the parolees from prior years would have been in the program might be impossible. In addition, there may be serious deficits in the information that can be gathered about the outcomes of these past parolees, because records of these outcomes might not have been kept. (We consider further the problems of using standard records in evaluation research in Section III.)

One might attempt to see some effect of the program by looking at all past and all present parolees and by using only the outcome data that are routinely recorded, but, in analogy to listening to a

radio program with a lot of static, this would increase the "noise" in the evaluation to such an extent that the "signal" of a true effect of the program might well be lost. Suppose, for example, that the job skills program were provided to 25 percent of a parolee population, and a before-after design on the institution were conducted. If recidivism rates of 40 percent for the "before" group and 35 percent for the "after" group were found, this *might* be the consequence of the program reducing recidivism from 40 percent to 20 percent among the parolees who participated in the program. But the overall change, from 40 percent to 35 percent, which is all that the researcher can actually observe, might be within the range of normal fluctuation from year to year (that is, recidivism averages 40 percent in the long run, but fluctuates normally between 35 percent and 45 percent in the short term). The "noise" of normal fluctuations in the larger population can thus be indistinguishable from a dramatic effect on the subgroup participating in the program.

Before-After Comparison Group Design

The quasi-experimental designs considered to this point are so susceptible to ambiguity that they are generally regarded by methodologists as useful only in a very limited set of program evaluations. They do, however, form the basis of more complex quasi-experimental designs that overcome some of the problems mentioned above. The greater complexity of the designs to which we now turn often results from the use of elements of more than one of the designs just discussed in order to use the strengths of one design to overcome the weaknesses of another.

We noted in our discussion of the simple comparison group design that one of the most potent threats to inference from that type of study is the likelihood that preexisting differences between the treatment and comparison groups cause differences in the outcomes being studied. Some of the ambiguity that would plague the results of such a study could be removed by adding observations of both the treatment and comparison groups prior to the beginning of the program, in the same fashion as in the before-after designs.

To use a before-after comparison group design to evaluate the job skills program, one would need to identify a comparison group in a similar prison and to record for both the treatment and comparison groups whether the inmate was regularly employed, for example, eight months prior to the present conviction and whether he or she is employed eight months after parole. Suppose the results show that 25 percent of the treatment group and 30 percent of the comparison group were employed at the time of the "before" observa-

tion and that 40 percent of the treatment group and 32 percent of the comparison group were employed at the time of the "after" observation. An inference that the program is effective in causing increased likelihood of employment might be made on the basis of the greater increase in employment in the treatment group. Because both groups are seeking postrelease employment in the same general economic climate, and because the comparison group did not increase as much as did the treatment group, the rival hypothesis that the increase is due to nationwide economic conditions is not viable, as it might have been if only a before-after design had been used. Because "before" observations are available on both groups, there is no need to wonder whether the comparison group had a different likelihood of employment prior to the study, a possibility that could not be ruled out if a simple comparison group design, with only "after" observations, had been used. In the present design, the direct observation of change eliminates some of the ambiguities that plagued the simple comparison group design.

BEFORE-AFTER
COMPARISON GROUP DESIGN

O_1	P	O_2
.....		
O_1		O_2

.....		indicates two groups (*not* constructed by random assignment)
O_1	=	the first observation of each group
P	=	exposure of one group to the program
O_2	=	the second observation of each group

Some sources of potential ambiguity still remain, however. It might be that local changes in job markets in the areas to which most of the individuals in the study are released could have caused the changes. If there were a local recession in the area to which most comparison group parolees were released and no such condition in the area to which most treatment group parolees were released, these results might occur whether the program was effective or not. It could be also that the comparison group, notwithstanding its previous higher employment rate, differs from the

treatment group in some way that causes the results. For example, the comparison group might contain a larger percentage of older persons, who might have already benefited from whatever tendency there is for better employment prospects with greater maturity, while the treatment group, with its younger members, may be benefiting simply by having become more mature and reaching some optimum employment age while incarcerated. There is also the possibility that other programs that are different in the two prisons could account for the results. In addition to these logical threats to unambiguous evaluation, there remains the problem of identifying the proper comparison group, which is as serious in this design as it is for the simple comparison group design. As noted above, it might be very difficult to know which soon-to-be-paroled inmates at the comparison prison would be eligible for the program if it were instituted there.

In our earlier discussion of the problems associated with the before-after design on individuals, we noted that an illusory effect might appear if the program were made available to an extremely needy group, because, in essence, the participants would have "nowhere to go but up" and they might seem to improve if a few had simply had worse luck at the time of the "before" observation than at the time of the "after" observation. The same source of ambiguity is present in before-after comparison group designs that select the treatment group on the basis of need and that use those deemed ineligible as the comparison group. Because natural variation can only lead to increases in the treatment group but can lead to either increases or decreases in the comparison group, this particular application of the present design can often lead to "pseudo-effects" that make the program appear effective when it in fact is not. When, in contrast, the treatment group is selected on the basis of high, rather than low, scores on the "before" measurement, the same process can work to make the program appear harmful when there is in fact no effect.

In general, the before-after comparison group design removes some, but not all, of the threats to unambiguous inference that exist in the simpler quasi-experimental designs. Uncertainty about preexisting differences on the characteristics thought to be affected by the program and uncertainty about the possibility that the results are due to some general time-based change are reduced or eliminated. Uncertainty about the potential effects of preexisting differences on characteristics that are not observed in the "before" observations, uncertainty about the potential effects of time-based changes that might affect one group but not the other, and the often critical problem of identifying an appropriate comparison group remain. Thus, although the before-after comparison group

design has moved closer to the randomized experiment in terms of eliminating ambiguity than the previously discussed quasi-experiments, there remain some important differences that point to the greater rigor of randomized experiments. However, when these uncertainties are judged not likely to have much force and when a good comparison group can be identified and observed, the before-after comparison group design is an alternative well worth considering.

Simple Time-Series Design

A popular quasi-experimental design that might be used to evaluate the job skills program is the "simple time-series design." One might observe for several years the recidivism rate of paroled inmates who would be eligible for the program if it existed, then institute the program and observe the recidivism rates of the participants during several years of operation of the program. If recidivism rates were constant or increased or decreased at a steady rate during the several years prior to the start of the program and if a sudden drop were seen when the program began and then the steady pattern continued at a lower level, the inference that the program reduced recidivism might be reasonable. For example, if 60 percent, 55 percent, and 50 percent, respectively, of the inmates released during the three years prior to the program recidivated within eighteen months of release and if 35 percent, 30 percent, and 25 percent, respectively, of the inmates released after the program commenced recidivated within eighteen months of release, there would be strong evidence that the program was effective.

The simple time-series design is really just an elaboration of the before-after design on institutions using multiple "before" and "after" observations. In the diagram, the vertical lines indicate different groups of similar individuals or cases processed before or after the program is instituted; the number of observations is arbitrary, subject to the requirements of the statistical procedures used

SIMPLE TIME-SERIES DESIGN

O_1	O_2	O_3	P O_4	P O_5	P O_6

⋮ indicates six groups studied (separated in time; *not* constructed by random assignment)

O_1, O_2, O_3 = observation of the first three groups prior to the start of the program

P = exposure of the second three groups to the program

O_4, O_5, O_6 = observation of the second three groups following exposure to the program

to detect time-based trends and separate these from the effect of the program.

The greatest benefit of the simple time-series design is that, unlike designs employing single "before" and "after" observations, it permits the identification of some time-based changes that might affect the characteristics under study and allows the consequences of these changes to be removed from the effects that might be attributed to the program. If only a single "before" and a single "after" observation had been made, it would be uncertain whether the drop from a 50 percent to a 35 percent recidivism rate were due to the program or to some general time-based change. The multiple "before" observations used in the time-series design, in contrast, make it clear that the change in recidivism rate is considerably more than would be expected from the general trend toward lower rates of recidivism. The multiple "after" observations make it clear that the drop in recidivism is not a temporary phenomenon.

Some problems still exist, however. The simple time-series design does not eliminate the possibility that some time-based change, starting at about the same time as the program and continuing through the remainder of the study, has caused the apparent effect. For example, suppose that one offense category could account for much of the recidivism of participants and potential participants in the program, and suppose that this behavior is decriminalized at about the same time the program is instituted. The results on recidivism of this one-time change would be the same as the results of an effective program. In addition, because it attempts to identify and rule out as rival hypotheses general temporal trends in the characteristics under study, the simple time-series design is likely to give ambiguous results if the characteristics vary so irregularly that there is no constancy to be found. Consider the difficulty in interpreting the results of the job skills program if, instead of showing the regular trends posited earlier, the recidivism rates were 60 percent, 20 percent, and 50 percent for the three years prior to the start of the program and 35 percent, 50 percent, and 25 percent for the three years subsequent to the start of the program. Such results would render ambiguous any interpretation of a time-series study.

The simple time-series design may require considerable delay in the testing of the program to allow for identification of similar individuals, observation of their outcomes, and collection of data for the multiple "before" observations. To use the full panoply of statistical methods for the analysis of time-series designs, it is often necessary to have observations on twenty-five or thirty time-separated groups. It is sometimes possible to use standard administra-

tive records from past years for this purpose, but this approach to time-series studies may lead to problems. Standard records are not designed to have the sensitivity required for high quality evaluation and may not contain enough information for identification of a comparison group or for measurement of characteristics or behavior that are relevant to the policy questions. If an attempt is made to generate a time-series design for evaluating the job skills program using standard records, there may be problems in deciding, on the basis of incomplete information, who would have received the training had it been available earlier and in determining whether these individuals found employment or recidivated. A common practical problem in time-series research arises from the likelihood that the program being studied is sometimes instituted when events have pointed to substantial problems in an institution, and these problems may have already led to both a deterioration of the quality of records and to multiple changes in policy and programs. In such an environment, both the logic and the practice of time-series research are threatened.

Despite its potential weaknesses, the simple time-series design is often a very powerful tool for evaluation; it is quite often a much stronger design than any of the quasi-experimental designs discussed above. If regular trends are present in the characteristics to be affected by the program and if records have been kept that are sufficiently detailed to allow immediate commencement of the study, this design can provide an efficient and relatively reliable approach to the study of justice system programs. The conditionals in the last sentence may pose insurmountable barriers in many instances, however, and these, together with the danger that a one-time event, unrelated to the program but occurring at the same time it is introduced, could produce error in inference, lead to the overall assessment that the randomized experiment is a more rigorous and more generally applicable design.

There are additional quasi-experimental designs that are elaborations of the simple designs already presented. For example, it is sometimes possible to combine the comparison group and the time-series design by making a series of observations on both the treatment group and the comparison group. In the job skills program example, one might use simultaneous time-series at two prisons, introducing the program at different times and using each prison as a comparison group for the other. This type of resourceful combination of quasi-experimental designs can often remove much of the ambiguity that would be inherent in the use of any single design, but it is the nature of quasi-experimental designs that there always remains some logical threat to unambiguous results. This, of course, is due to the existence of some systematic difference be-

tween the treatment and comparison groups or between the current and past situations.

Additional Control Procedures

The concern in research design is elimination of possible alternative explanations for the results of a study, with the ultimate, but perhaps unreachable, goal of leaving only the innovation itself as the cause for any effects observed. We have attempted to point out how observations before introduction of the program, comparison groups, and randomly constituted groups are employed to try to remove ambiguity from an evaluation. There are two additional techniques that can be used in either quasi-experimental or randomized experimental designs to further reduce the ambiguity of research results. These techniques are termed "matching" and "statistical control." It should be stressed, though, that these techniques can never raise a quasi-experimental design to the level of methodological rigor of a randomized experimental design.

A matching procedure, in its simplest form, requires the pairing of subjects who share characteristics that might influence the results of the study. One member of each matched pair is subjected to the program; the other is not. The matching technique attempts to assure that any subsequent differences between the "treated" and the "not treated" subjects cannot be attributed to the characteristics on which the matching is based, because those characteristics occur equally in each group of subjects. For example, in a comparison group evaluation of the job skills program, one might be concerned that inmates in a comparison group at another prison might have educational backgrounds substantially lower than do inmates at the prison testing the program. Because one would expect education to affect postrelease job prospects, the difference between the comparison and treatment groups raises a strong rival hypothesis for any apparent benefit of the program. But if each inmate in the comparison group were paired with an inmate of similar education in the treatment group, and if only these matched subjects were used in the evaluation, the education-based ambiguity might be eliminated.

Matching can be employed in a randomized experiment if, for each matched pair of subjects, one subject is randomly assigned to the experimental group and the other to the control group. (Statisticians often use the term "blocking" to refer to random assignment of matched subjects.) The randomization insures against systematic differences in unmatched characteristics, while matching insures against *any* differences in the matched characteristics. This

may reduce the likelihood of random differences between the groups and thus make the evaluation more able to detect small differences caused by the program. Thus, if matched pairs of inmates were identified and a random procedure were used to assign one member of each pair to the job skills program, some of the natural variation between individuals would be eliminated (in the sense that it would be known and removed from consideration), and the statistical tests that are needed to confirm differences in outcome between the experimental and control groups might be more sensitive.

Statistical control through techniques such as "covariance adjustment" may be thought of as a sophisticated type of matching, in which statistical techniques are used to "predict" what outcome would be expected from the characteristics of the individual subject. Each subject is then "matched" with the predicted outcome for that subject. This technique uses statistical procedures to "adjust" for some of the differences between the subjects, and looks for effects that cannot be accounted for by this adjustment. Both matching and statistical adjustment techniques can increase the precision of randomized experiments by reducing the likelihood of the random difference problem referred to earlier.

These techniques are applicable to both quasi-experimental and randomized experimental designs, but in quasi-experimental settings the use of matching or statistical adjustment may result in the appearance of "pseudo-effects." These are differences that appear to be consequences of the treatment but that are in fact attributable to imperfections or irrelevance in the factors used in the matching or adjustment. As a result, these control techniques can sometimes suggest a program effect even when there was no such effect. This problem arises in quasi-experiments when the treatment and comparison groups in their unrefined condition would differ even in the absence of treatment effects. Because there is inaccuracy in the measure or classification used in matching or statistical adjustment procedures to "equate" the groups, the techniques lead to underadjustment of the true differences. The natural differences reappear on measures that are supposed to tap the program's effects, thus producing pseudo-effects. This is a very complex issue, referred to by methodologists as the problem of "error in variables," or "regression artifacts." We need not pursue it further, but the problem should be noted because it is sometimes tempting to see matching or statistical adjustment as "cures" that render quasi-experimental designs as powerful as randomized experimental designs. They are not, but are instead valuable—but tricky—adjuncts to the inherent logical advantages or disadvantages of a particular design for demonstrating causal relationships.

It should also be noted that this use of matching or statistical control in quasi-experiments is logically suspect on other grounds. There is always the possibility that some important factor has been omitted from the matching scheme. Moreover, in many program contexts the exact criteria for selection of treatment subjects cannot be determined (due to self-selection if, for example, the program is such that some treatment subjects can choose to participate) or replicated (due to absence of the relevant data for comparison subjects). Nor, of course, can these techniques adjust for differences that stem from the systematic difference that defines the two groups: if all treatment subjects and no comparison subjects are volunteers, we cannot "adjust away" results that may be explained on that basis.

III. MEASUREMENT

The topics just discussed concern construction of an evaluation study so that it will yield meaningful information from observation of the characteristics and behavior thought to be affected by the program. But some programs are intended to affect characteristics that are not observable, and no degree of rigor in design can make an empirical evaluation speak to the most important effects of such programs. As noted at the beginning of the previous section, a change from twelve-person to six-person juries, if motivated by a concern to increase the efficiency of jury trials without affecting the objective fairness of verdicts, cannot be fully evaluated by any empirical evaluation study because one cannot systematically observe or measure the objective fairness of verdicts. (One could, however, measure the extent to which litigants *perceived* the verdicts to be fair, but, as noted below, this is a different question.) Only those potential program effects that are amenable to general observation, measurement, or counting, or that have some indicia that can be observed, measured, or counted, can be subjected to empirical study. A particular program effect may not be open to evaluation and still be the prime consideration in policy decisions, of course, but it is simply outside the province of scientific evaluation.

Another aspect of measurement must be considered, even though it may appear trivial at first glance. It is important to remember that there is a difference between objective program effects and subjective perceptions of program effects. It is often possible to assess both objective and subjective reactions to a program, and often both are important to policy decisions. But it is dangerous to confuse the consequences of a program with what people involved in the program *think* its consequences are. Consider the difference

between conducting a rigorous evaluation of the objective consequences of the job skills program for postrelease employment and recidivism and conducting a survey of parolees to ask whether they think the program helped them. A positive result of the objective evaluation would offer concrete evidence of the effectiveness of the program. But a positive result of the subjective survey, while encouraging, would be open to a variety of interpretations. (It is a truism in evaluation research that perceptions of the effectiveness of programs by those involved in them are quite often positive, even when objective evaluations of the same programs show no positive effect.) Similarly, it would be dangerous to base an evaluation only on the impressions of those who administer the job skills program. A variety of psychological factors—for example, psychological investment in the success of the program—affect the beliefs of those involved in a program, and these factors can lead to impressions that do not reflect reality.

This is not to say that subjective reactions are not important issues in either the evaluation or the policy decisions that must ultimately be made. Careful measurement of subjective impressions can offer much to the interpretation of objective findings, and positive subjective reactions are often themselves goals of a program. Most evaluations should involve measurement of both objective and subjective factors. It is only the attempt to substitute one for the other that we caution against here. (Note that there is also danger in attempting to substitute objective effects for subjective reactions. To find that a program benefits participants, in that it achieves goals that the program designer thinks are the goals of the participants, is not necessarily evidence that the participants share the designer's goals or that they in fact think the program is effective.)

Assuming that the matters a program may affect can be measured at all, the practical concern becomes the choice of what particular characteristics are to be measured. For even the most rigorous evaluation to be useful, it must ask the right questions. This can be accomplished only if the evaluator is adequately informed about the theory, or rationale, of the program, and only if those by whose authority the program is to be instituted are willing to work with the evaluator to determine what observable characteristics will speak to the policy questions under consideration. This task may be difficult at times, because the broad issues of concern to the policy maker have to be transformed into quite specific characteristics upon which the evaluator may collect data. However, this collaboration is crucial to the success of any evaluation.

Virtually all measures or observations in empirical research are subject to some "error," and a major part of the evaluation effort is to find or construct measures for which such error is small and not

threatening to the overall accuracy of the study. We hasten to point out that the term "error" as used by methodologists includes variability that is irrelevant to the study and that has no worse effect than increasing the "noise" surrounding the program's effects. Methodologists distinguish between two general types of error in measurement, errors in the "validity" of the measure and errors in the "reliability" of the measure. We offer brief descriptions of each type of error below to convey the concerns that an evaluator will have in deciding how to measure the effects of a program.

The characteristics chosen to be measured in order to ascertain whether they are affected by an experimental program will often be surrogates for the matter which is of important policy relevance. Such measures are of greater or lesser "validity" depending upon how well they reflect the matter of genuine interest. For example, the goal of the job skills program is to reduce recidivism—the incidence of crimes committed by parolees—and we may choose to measure this by collecting information on new crimes of which the subject is convicted. But convictions are clearly less frequent than actual acts of crime, and conviction for an act may be affected by factors that do not affect occurrence of the act. Thus, the incidence of convictions may not be influenced by the experimental program although the incidence of actual acts of recidivism is affected. The validity of convictions as a measure of recidivism may be rather weak when applied to this particular evaluation, and it might be that the incidence of arrest, regardless of subsequent conviction, is a more valid and sensitive index of recidivism for purposes of the research. In general, the objective is to construct measures that are affected by the same factors as are the characteristics they index. Invalid measures can seriously threaten the accuracy of an evaluation, because they can lead to mistaken impressions about what the results of the study actually mean.

The "reliability" of a measure has to do with its consistency from case to case and time to time. For example, if regular, but part-time employment were counted as a job for some parolees who had the job skills training but not for others who had the training, the reliability of the employment data would be reduced. Similarly, if there are many errors in the standard records on which a time-series study is based, the study may suffer from the unreliability of the measures taken from those records.

Unreliable measures generally pose a less serious threat to an evaluation study than do invalid measures. Invalid measures can lead to the more dangerous error of incorrect interpretation of what an observed effect means, while unreliable measures may simply conceal that an effect has occurred or exaggerate its magnitude. Of course, every effort should be made to construct measures that are both valid and reliable.

A final issue in measurement concerns the use of standard administrative records as sources of data for an evaluation. Data from such records are often an essential part of evaluation studies, especially when time-series designs are used. It is necessary, however, to exercise caution when using administrative records. Administrative records are designed for purposes other than evaluation, and they sometimes do not contain the information that is necessary for the evaluation. To rely too heavily on data from such records is to risk an evaluation that addresses what is in the record rather than what would best inform the policy decisions that must be made. In addition, many large record systems suffer from reliability and validity problems to such an extent that they are of limited value for sensitive evaluation research. Errors in the recording or coding of information contribute to reliability problems by introducing "noise" in the evaluation data. Such practices as frequent, unpublished changes in the definition of recorded entries contribute to validity problems by raising the possibility that the evaluator will be mistaken about what events or acts are implied by the entries in the records. Improvements in the quality and consistency of administrative records may help alleviate some of these problems, and consideration, by those who design and keep such records, of their potential usefulness for evaluation may render the data from them more valuable for researchers. But until this occurs, the problem remains.

IV. INTERPRETATION OF RESULTS

We noted above that there must be close collaboration between the policy maker and the evaluator if an evaluation study is to pose the proper questions for later policy decisions. Only if the evaluator is familiar with the theory of the program can the specific data collected speak to whether the program has the effects it is designed to have. Similarly, only if there is sufficient collaboration between the evaluator and those who administer the experimental program can the evaluator offer realistic interpretations of the basic findings of the research. The evaluator must have an accurate conception of the practice of the program, as well as an accurate conception of its theory, if the proper interpretation is to be found. If possible, an evaluation study should collect information on the day-to-day practice of the program, but such information will seldom be all that is needed for a good evaluation. For example, an evaluation of the job skills program might include repeated

testing of program participants to determine whether job skills are indeed being improved or whether any apparent effect is due to some other factor in the program (for example, whether beneficial outcomes are resulting from simply paying more attention to the inmates). And it is crucial that the evaluator be informed directly of how the program is put into practice so that rival hypotheses can be considered and tested and so that the evaluation can offer information not only on whether the program works but also on how it works.

Our earlier caution about using the impressions of program providers in place of rigorous evaluation of the objective consequences of a program should not be taken to mean that such impressions cannot be very useful to the overall evaluation effort. Impressions that a program is producing the benefits it is intended to produce do not prove this to be the case, but impressions about the processes involved in the workings of the program can provide valuable clues to where the evaluator should look for objective data about potential problems and accomplishments of the program. Thus, if those who teach the skills in our hypothetical job skills program say that the time allotted for the training is too brief, it does not necessarily mean that the program is ineffective, but it should alert the evaluator to the need for additional data collection on, for example, whether program participants must learn additional skills before they can make full use of those taught in the program.

In general, there will be problems of ambiguity in the evaluation, whatever design and measurement methods are used, if the evaluator cannot determine how the findings of the experiment relate to what would be seen in the full-scale application of the program as general policy. Knowledge of the practice of the program is necessary for this determination. If the practice departs from the theory, it is uncertain that the same results would obtain if new programs follow the theory, but not the practice employed in the experiment. Another potential ambiguity that can plague efforts to extend the findings of the research to the situation that will exist when the program is no longer experimental is the possibility that the results of the evaluation were affected by a phenomenon termed "reactivity," an issue that we now consider.

Social scientists have long known that the very act of studying human beings can cause them to act in ways other than they normally would. The knowledge that one is involved in an experimental program, that one's behavior is being observed and recorded, or that one has been placed in a program on the basis of random assignment can sometimes lead to responses that would not occur if the program were in routine use, if no special observations were being made, or if assignments were based on characteristics of the

subjects. Such behavior is said to be "reactive." In the hypothetical job skills program, reactivity might occur if those involved in the study knew that someone was monitoring closely whether or not they were employed and if they therefore made special efforts to find employment. Given such knowledge on the part of those being studied, the results of the study might be different from the results that would actually occur if the program were not under study and employment were not monitored beyond what is standard for parolees. Another example of reactivity was mentioned earlier: if inmates in a control or comparison group know of the job skills program and resent not receiving it, this resentment may lead to behavior that is not truly characteristic of the status quo situation.

It is, of course, desirable to minimize the likelihood that behavior observed in an evaluation study is affected by reactivity. Validity of the results of the experiment requires that the responses of subjects exposed to the experimental treatment be as much as possible like those of subjects who might in the future receive the treatment on a routine basis and that the responses of subjects used for comparison be as much as possible like those of subjects who would not receive the program if it is abandoned in the future. One means of attempting to avoid reactivity is to misinform or not inform the subjects about aspects of the experiment that might cause reactivity. In a randomized experiment, one might fear that either control or experimental group subjects will react to the random assignment with behavior that they would not show otherwise, and one might therefore avoid telling the subjects that they had been randomly assigned. Or one might fear that subjects will react to the intense observation needed to assess effects of the program, and one might therefore avoid telling the subjects that they are being observed or that data are being collected on them. Of course, there is always the possibility that the deception will be discovered and that the subjects will be even more reactive to knowledge of the deception than they would have been to knowledge of the design or the observation.

This appendix raises the problem of reactivity and its possible solution by deception not to encourage the use of deception, but only because it is an issue that sometimes arises in evaluation research. The body of the report discusses the issue and its ethical implications, and our concern here is simply to alert the reader to the reasons that might prompt one to consider the use of deception.

V. TECHNIQUES FOR MAINTAINING PRIVACY AND CONFIDENTIALITY

Another issue that often arises in program evaluation is protecting the privacy of individual subjects and the confidentiality of information pertaining to them. A number of methods have recently been devised to allow researchers to obtain and use information while providing such protection.[40] All of these methods attempt to limit the capacity to attribute sensitive characteristics to an individual, while allowing analysis of the characteristics of the group to which the individual belongs. These techniques can be divided into two broad categories—procedural methods that permit record linkage, and statistical methods that permit the collection and holding of sensitive information.

Procedural Solutions to Obtaining Data from Restricted Records

Frequently, a program evaluation can be facilitated by information in confidential records that have been constructed for other research or administrative purposes. For example, in studying the impact of the hypothetical job skills program, it would be helpful to follow the earnings history of the former inmates for several years after their participation in the program and release into the community. Even if the participants in the program agree to continue providing information, the passage of time would probably result in great difficulty in collecting accurate employment data, because of the practical problems of maintaining contact and cooperation over long periods. Another option is to use the record of earnings maintained by the Social Security Administration or the Internal Revenue Service. This would permit the collection of accurate data over a long period of time with little attrition from the study. Often, however, access to such information is restricted by assurances of confidentiality or statutory protection.

One procedural solution to such a problem is to combine individual data into small groups and analyze the groups as though they were individuals. First the researcher constructs small clusters of three or more individuals within each of the general experimental groups. The identification of the individuals within each cluster is then sent to the government agency or archive maintaining the employment records. The archive locates its records for each individual in a cluster, computes average reported earnings for the

40. For further discussion of these techniques the reader is directed to R. Boruch & J. Cecil, Assuring the Confidentiality of Social Research Data (1979).

cluster, then links that information to the cluster records sent by the researcher. All individual identification is then removed from the records and the anonymous data are returned to the researcher for analysis. The result is a data set that links archive information to the information collected by the researcher without breaching the privacy obligations of the archive or the confidentiality assurances of the researcher.

Such a technique permits research access to a great variety of restricted data archives, such as bank records, employment records, Internal Revenue Service records, and school files. Many variations on this strategy have been developed. However, care must be taken to aggregate data in such a way that it will not be possible for the archivist or the researcher to deduce information about a single individual from the statistical data describing the cluster.

Other procedural means exist for increasing the confidentiality of data, including purging of identifying information as soon as it has served its purpose, or, if such information must be retained, separating the data into sets and distributing them among several persons in a way that prevents any individual researcher from knowing both the identifying information and the data it links to particular individuals. For example, in order to isolate subject identification from questionnaire responses in a long-term study of criminal behavior, identifying information and responses can be linked by code numbers. The researcher possesses the responses and associated code numbers, while a trustee possesses the subjects' names and addresses and the code numbers. Follow-up questionnaires would be mailed by the trustee when the researcher sends him or her the code numbers of the subjects to be surveyed, and the completed questionnaires, identified only by code number, would be returned to the researcher. Even more secure, and complex, schemes can be used when extreme caution is required.

Statistical Means of Maintaining Privacy and Confidentiality

The hypothetical job training program is also intended to reduce subsequent criminal behavior. One direct means of gaining this information is to ask the former participant in an interview how frequently he or she has engaged in criminal behavior in recent weeks. Such an approach obviously encounters a number of problems. The participant may be reluctant to share such information with the researcher, despite assurances that the information will remain confidential. The researcher also may be reluctant to collect such information, because it may expose research participants to increased risk of prosecution.

Several statistical methods have been developed to minimize such problems. In general, the statistical methods introduce a known amount of error into an individual response, making it impossible to deduce the individual's answer but still permitting conclusions about the group to which the individual belongs. One of the most common statistical approaches is known as the "randomized response method." In terms introduced in the discussion of measurement above, this method introduces sufficient unreliability into the data to make them useless for any purpose other than the aggregate analysis to be used in the research. When used in surveys, these procedures can actually improve the validity of the data, because greater candor can be expected. In a simple version of this approach, the researcher presents each respondent with two questions, one innocuous and one sensitive, such as, "Did you buy a newspaper yesterday?" and "Did you participate in criminal behavior within the past week?" Each question must be answerable with a "yes" or "no" response. Before answering, the respondent is asked to roll a die out of sight of the interviewer and to answer the first, innocuous, question if a one, two, three, four, or five turns up on the die, and to answer the second, sensitive, question if a six turns up. Because the interviewer does not observe the roll of the die, only the respondent will know which of the two questions is being answered. However, given a proper sampling scheme and the odds of answering each question, it is possible to estimate statistically the proportion of persons who answered "yes" to the sensitive question without knowing the true response of any individual respondent. It would be possible, for example, to determine what proportion of the group of respondents had engaged in criminal activity within the past week without determining the true level of criminal activity of any of the individual respondents. These methods have been used by researchers to examine criminal behavior, sexual behavior, and racist attitudes.

If confidentiality in the data record, rather than privacy in the response itself, is the primary concern, it is possible to use similar techniques after the data have been collected. Thus, a researcher might randomly change a percentage of the data records of sensitive information. Again, the basic concept is to introduce random, and thus statistically tractable, error that renders the data usable for the research but unusable for any purpose relating to the individual subjects.

These methods solve some of the problems that arise from concerns over privacy and confidentiality, but they have notable disadvantages. Large samples are usually required, making the research more expensive. They require a measure of technical sophistication, and consequently increase the complexity of the research. If

the evaluation turns on questions that can only be answered by collecting sensitive information or by obtaining data from archives to which access is normally restricted, however, the methods described here are worthy of consideration.

Bibliography

Boruch, R.F., & Cecil, J.S. *Assuring the Confidentiality of Social Research Data*. University of Pennsylvania Press, 1979.

Campbell, D.T., & Stanley, J.C. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, 1966.

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

Rossi, P.H., & Freeman, H.E. *Evaluation: A Systematic Approach*. Sage, 1979.

INDEX

C

D

E

N

O

P

* U.S. GOVERNMENT PRINTING OFFICE: 1981 O - 349-101