Handbook of Resources for Criminal Justice Evaluators

U.S. Department of Justice
Law Enforcement Assistance Administration
National Institute of Law Enforcement and Criminal Justice

Handbook of Resources for Criminal Justice Evaluators

Anne L. Schneider, Ph.D.
Principal Investigator
Peter R. Schneider, Ph.D.
Co-Principal Investigator
L. A. Wilson II, Ph.D.
Research Scientist
William R. Griffith, M.A.
Research Assistant
Jerry F. Medler, Ph.D.
Research Scientist
Howard F. Feinman, J.D.
Research Scientist

August 1978

U.S. Department of Justice
Law Enforcement Assistance Administration
National Institute of Law Enforcement and Criminal Justice

Law Enforcement Assistance Administration
Henry S. Dogin
Administrator
Homer F. Broome, Jr.
Deputy Administrator for Administration
National Institute of
Law Enforcement and Criminal Justice
Harry M. Bratt
Acting Director

Reprinted by the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice in cooperation with the State of Washington Office of Financial Management, Law and Justice Planning Division.

This project was supported by Grant Number DF-1540, awarded to the Institute of Policy Analysis by the State of Washington Office of Financial Management, Law and Justice Planning Division. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

PREFACE

The Handbook of Resources for Evaluators was prepared as part of the LEAA-funded Model Evaluation Program in the State of Washington. Evaluators who participated in that program will recognize some of the materials, but others were prepared especially for the Handbook. Much of the written work and verbal presentations prepared during the course of the Model Evaluation Program have been updated, expanded, and revised for inclusion in the Handbook.

The major purpose of the Handbook is to provide evaluators, planners, and decision makers with information about techniques that could be used to overcome the more typical problems encountered in criminal justice evaluation. The techniques described and discussed in the Handbook address four different kinds of problems:

First, there are technical and research-related problems which, if not overcome, result in evaluations that do not contain valid, accurate answers to the questions that the evaluation sought to answer.

The second problem concerns what type of evaluation should be done and what questions should be answered by it. Related to this is the problem of delineating the roles of the evaluator, the planner, the project director, and other decision makers in determining the type of evaluation, the questions to be answered, and the degree of confidence needed in the conclusions that are drawn. An evaluation that does not provide valid answers to relevant questions is not likely to be used by anyone.

Third, and often overlooked, are the problems evaluators and others

must face in insuring the privacy and confidentiality of data collected and the protection of human subjects when they are affected by the evaluation study.

The fourth problem is that evaluators do not tend to stay in their jobs for particularly long periods of time and there is insufficient documentation available to newly hired evaluators about local resources, organizational structure of the office, administrative procedures, and sources of data.

Thus, the Handbook is expected to be useful to current evaluators, new evaluators, and to planners, project directors, and other decision makers in their efforts to produce evaluations that contain scientifically valid information that will be useful in the decision making process.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

*This is one of the two sections omitted from the copies distributed by LEAA because it pertains only to the State of Washington.

PART I

TECHNIQUES FOR OVERCOMING COMMON PROBLEMS IN CRIMINAL JUSTICE EVALUATION

INTRODUCTION TO PART I

The purpose of Part I is to describe the major problems in criminal justice evaluation and to suggest alternative procedures for overcoming these problems.

Section 1 delineates the general categories of problems in criminal justice evaluation and reviews the techniques suggested in the papers contained within Part I for dealing with those problems.

Section 2 contains nine short papers, each of which describes a particular research or technical procedure that could be used by evaluators in order to increase the validity and accuracy of conclusions drawn by the evaluation.

Section 3 is primarily for planners, project directors, and decision makers, but also would be of value to evaluators. The six papers in this section examine techniques that could be used to determine the type of evaluation that should be conducted, the questions that should be addressed, and/or techniques for integrating the role of evaluators with the roles of planners, project directors, and other decision makers.

Section 4 contains eight evaluations (or excerpts from evaluations) of projects within the State of Washington. Each was selected because it demonstrates an innovative or exemplary approach for overcoming problems in field evaluation or because it applies some of the principles discussed in the other papers in Part I.

Section 5 contains the most recent LEAA regulations about privacy, confidentiality, security, and protection of human subjects. In addition, this section summarizes the types of forms that are needed, provides actual examples of procedures that were used, and discusses the issues involved in obtaining informed consent.

SECTION 1

INTRODUCTION AND SYNTHESIS

Abstract

The introduction seeks to pull together the general categories of problems and the approaches suggested for overcoming them. Although the discussion of each problem-solving technique is quite brief, it is sufficient to illustrate the basic rationale of the technique and how (or when) it could be used.

INTRODUCTION

The purpose of evaluation is to produce scientifically valid information or conclusions that will be useful within the planning and decision making processes. In order to obtain valid information, evaluators must recognize the inherent problems in field research and must learn to use the techniques that are available for overcoming them. To produce useful information the evaluator must apply the research techniques to the questions or propositions that will yield the most relevant information for the persons who are expected to utilize the results.

Planners, project directors, and decision makers also have a critical role in whether evaluation findings will be valid and useful. The integration of the work done by evaluators with that done by planners, project directors, and other decision makers begins with a common understanding of the types of evaluation that are available, the techniques for identifying important questions to be addressed, and the procedures that must be followed by the project if the evaluator is to be successful in efforts to produce valid as well as useful information.

The materials presented in Section 2 are mainly of interest to evaluators, since these deal with alternative techniques that could be used to overcome the more common types of research-related problems in evaluation. The materials in Section 3 are designated primarily for planners, project directors, or other decision makers, but should be studied carefully by evaluators since these provide a common

framework for identifying the types of evaluation, the questions to be addressed, and the involvement of the evaluator in project planning, project implementation, and project operation.

TECHNIQUES FOR STRENGTHENING EVALUATION DESIGNS

The most common technical weakness in evaluation research is the use of an evaluation design that is too weak to rule out alternative explanations (threats to validity) for the observed or apparent effects of the project. The first step in overcoming this problem is for evaluators to be more conscious of the alternative explanations that may confound their conclusions. By anticipating confounding factors evaluators should be better able to select an appropriate design or to collect additional data that could be used to test whether certain alternative explanations have, in fact, been confounded with the apparent impact of the project on the problem it was designed to solve or ameliorate.

"A Review of Threats to Validity" provides a semi-technical description of the general categories of alternative explanations that most often plague evaluation research in criminal justice. The approach differs in two important ways from the more common explication of Campbell and Stanley's "threats to validity."[1] First, the discussion does not provide a simple "yes or no" answer to whether a particular design automatically "solves" a threat to validity problem. Instead, the conditions that are needed <u>if</u> a particular design can be generally relied upon to rule out the various alternative explanations are identified.

Second, the paper is organized so that each threat is described and discussed in relation to several relevant designs rather than having each type of <u>design</u> presented and then discussed in relation to each threat to validity. This approach should make it easier for the evaluator

(whose choice of design is often quite limited) to identify the relevant threats to validity and then to determine what could be done to overcome them within the constraints imposed by the situation.

Experimental designs are the most powerful that can be used for assessing the effectiveness of a project. Although rarely used in evaluation research, there is some evidence that when evaluators are involved in the planning and design of the project prior to its implementation they can be quite successful in obtaining true experimental conditions in the field setting.[2] One situation that is especially conducive for random assignment is when the evaluator can assist the planner or project director in identifying a pool of eligible persons (or cases or areas) who need the treatment (intervention) and, in addition, the resources for the project are less than those needed to handle all eligible clients (or cases or areas). In this situation random assignment is a fair and equitable way to distribute the limited services to those who need it. A second situation occurs when a project has several components or alternative strategies and tests the effectiveness of each. In the absence of any knowledge as to which strategy is best, random assignment is a fair method of determining who gets what within the project and will permit a rigorous test of which strategies are more effective.

When random assignment is not used the evaluator must rely on some type of quasi-experimental design. Two of the major categories of quasi-experimental designs are discussed in separate papers within the Handbook: Comparison group designs and interrupted time series designs.

In the paper *"Comparison Group Designs"* the logic of these designs is described and their usefulness under different field conditions is analyzed. The analysis in the paper leads to several important conclusions about the use of comparison groups.

First, the argument is made that comparison groups do not have to be perfectly equivalent to the treatment group in order to be useful. Instead, the comparison group should be equivalent to the treatment group in terms of the variables that are relevant to ruling out one or more specific threats to validity. Thus, with the use of several comparison groups--none of which is perfectly equivalent--the evaluator could, in some situations, rule out most or all of the alternative explanations for the observed effects of the project.

Second, the use of multiple regression is suggested as a more appropriate approach than actuarial tables or matched pairs.

Third, the paper points out that the selection procedures of the treatment group are critical for ascertaining whether a comparison group will be effective in ruling out alternative explanations. When these procedures are based on quantitative eligibility rules, comparison groups will be much more useful than they are when judgmental decisions are made. Judgmental decisions that place the "easy" cases into one group and the "hard" cases into another are the most difficult to handle. In the latter situation none of the procedures for obtaining comparison groups or analyzing data (i.e., matched pairs, actuarial tables, and multiple regression) necessarily will permit valid conclusions to be drawn. Nevertheless, a comparison group approach is recommended (over a pre-post design) because it is more likely that this

design can be used to rule out the alternative explanations.

"Interrupted Time Series Designs" are described in the third paper. These are among the strongest that can be used in field evaluation and often place so few constraints on project operation that they are more feasible and practical than any other approach available to the evaluator. The discussion in the Handbook covers six topics about interrupted time series: The logic of interrupted time series, the patterns of change that are of interest, the different types of interrupted time series designs, the conditions under which interrupted time series (depending on the particular type) controls for each of the major threats to validity, the statistical procedures available for analyzing interrupted time series data, and techniques for handling the problem of autocorrelation in the residuals.

Distinctions are made among the following interrupted time series designs:

1. Ordinary Interrupted Time Series: A series of pre and post measures on the same group or area.

2. Different Group Interrupted Time Series: A series of pre and post measures with the pre-project measures on groups that would have been in the treatment if the project had existed at that time and the post measures on persons actually in the treatment program.

3. Multiple Interrupted Time Series: A series of pre and post measures for the treatment group or area and a series of pre and post measures for a comparison group.

4. Experimental Interrupted Time Series: A series of pre and post measures on randomly selected treatment and control groups.

5. Individual-Level Interrupted Time Series: Measures on a series of pre-project persons (rather than groups) who entered the system at several points in time prior to the project and a series of observations on individuals who entered the system at several points in time after the project was implemented.

Regardless of which design is used, the purpose of interrupted time series is to identify changes in the trend that could be attributable to the project intervention and/or changes in the level. Interrupted time series designs clearly are superior to pre-post designs and in many situations will be more useful than comparison group designs for ruling out alternative explanations. The strength of interrupted time series, however, depends on which of the designs is used and the nature of the threats to validity. Maturation, regression effects, and testing effects can be serious problems for the design (when observations are on the same group or area), but the different group design generally would be effective in handling problems introduced by maturation effects or testing effects. A multiple time series in which the evaluator has both a treatment and a comparison group, with pre and post observations on each, is especially powerful.

Statistical procedures for analyzing interrupted time series data are rather complex and present a number of problems for the evaluator. The paper describes many of the statistical procedures and notes that six of them (Walker-Lev, mood, double mood, analysis of covariance, Chow test, and dummy variable regression analysis) are all based on linear regression approaches. The Walker-Lev, ANCOVA, and dummy variable regression are identical to each other. The problem of autocrrelation in the residuals is explained and procedures that could be used by the evaluator to resolve this problem are presented.

ANALYSIS AND STATISTICAL TECHNIQUES FOR QUASI-EXPERIMENTAL DESIGNS

The statistical and analytical procedures for experimental research are usually quite straightforward and are among those taught in the most basic university statistics courses. In contrast, the analysis of data from quasi-experimental designs is considerably more complex. As indicated in the last section of the interrupted time series paper, one generally will find many statistics that can be used. Some of these are identical to one another but have different names because they were developed within different academic disciplines. The statistics and analysis procedures described within the Handbook for use with quasi-experimental designs have one factor in common: Each utilizes some technique to control or "hold constant" the variables other than the treatment of interest so that the "true" impact of the project (or other independent variable of interest) can be separated from the confounding effects of the other variables.

The paper *"The Intuitive Logic of Multiple Regression Analysis"* provides a step by step explication of the logic underlying regression/correlation approaches and a step by step description of how errors on hypothetical cases are manipulated to provide the summary statistics from correlation/regression analysis. Substantive interpretations are provided that illustrate the differences in meaning for the correlation coefficient (r), the coefficient of determination (r^2), the regression coefficient (b), and the intercept (α). In addition to the general description, several of the key assumptions of regression analysis are

explained in non-technical terms and the impact of violating these assumptions on the results of the analysis is discussed.

The paper *"Prediction Models"* covers three distinct topics: actuarial tables, multiple classification analysis, and linear prediction (regression analysis). All three approaches could be (or have been) used for the purpose of developing predictions of what the scores of project clients or areas on the dependent variable (performance measure) would have been if the project had not existed. Thus, these are analytical and statistical techniques that are especially appropriate when the evaluator has a non-randomly selected comparison group (concurrent or historical) and has individual-level data about the persons in both the treatment and the comparison groups. If the methodology produces a reliable and valid estimate of what the scores would have been in the absence of the treatment, then the techniques would be a substitute for experimental designs. Unfortunately, none of the procedures is a completely reliable substitute for random assignment, but these are powerful tools that, in the proper situations, would permit the evaluator to draw a valid conclusion that otherwise would have been impossible.

The paper *"Applications of ARIMA and ANCOVA to Interrupted Time Series"* is a rather technical presentation of the fundamental differences (and similarities) between these two. In the ANCOVA approach, time (measured in months, years, weeks, or other similar units) is an independent variable in the regression equation. A linear prediction or projection is made from the pre-project observations into the post-project time period in order to ascertain changes either in the trend or the level of the series. The ARIMA models do not use time in the

equation, but instead base future predictions of the values of observations on the immediate past value of the dependent variable or on patterns in the errors of previous predictions. Although arguments can be made that ARIMA models are more appropriate for social science data than the linear trend predictions used in ANCOVA, the former are more difficult to use because well-documented and well-developed statistical routines are not generally available to most evaluators.

Regardless of the design or analysis procedures that are used, the estimate of whether the apparent impact of the project is due to chance must be ascertained. Whether an apparent impact will be statistically significant at a particular level (such as .05) depends on the magnitude of the impact and on the size of the sample. It would be a rather embarrassing situation for an evaluator to draw a sample for the evaluation which was so small that even if the project achieved its quantitative goal (of reducing crime, for example, by 10 percent) the results would not be statistically significant at the .05 level. In order to avoid that problem, a paper *"Determining Appropriate Sample Size"* is included in the Handbook. The paper provides tables that show the size of samples needed in order for specific differences in proportion to be statistically significant at the .05 and .01 levels.

TECHNIQUES FOR MINIMIZING MEASUREMENT PROBLEMS

Another major problem in evaluation research involves the reliability and validity of data. In *"An Introduction to Measurement Problems"* reliability and validity are defined and discussed within the context of criminal justice evaluation. Of particular importance is the fact that some reliability and validity problems result in a reversal of the true direction of the relationship and, for example, make it appear that the treatment group had higher recidivism rates than the comparison group when, in fact, the opposite was true. Other types of error will not affect the true direction of the relationship, but will depress the values of statistics used to test the significance of the differences. This, in turn, leads the evaluator to conclude that the project was not effective when, in fact, it was. Techniques for identifying reliability and validity problems are discussed in the paper, as are procedures that the evaluator could use to interpret the data after ascertaining the nature of the reliability and validity problems.

The paper *"Measuring Change in the Crime Rate"* identifies three basic sets of data that could be used: official crime statistics (reported crime), two or more victimization surveys taken at different points in time, and one victimization survey divided by the months included within the recall period. The types of errors and problems with each of these procedures is discussed in this paper. In general, the official statistics would be better unless there are reasons to believe

that the project altered the inclination of victims to report crimes to the authorities. If so, then the changes in officially reported crime rates will reflect not only changes in the frequency of offenses, but also changes in reporting.

The use of two victimization surveys (one pre and one post) is confounded by problems of comparing two surveys unless they were conducted under virtually identical conditions. Furthermore, unless the evaluator has developed a rather elaborate sampling plan to include some "treated" households and some "untreated" households, the results of the two surveys will be one of the weakest of all types of comparisons--one pre and one post observation for the entire geographical area.

A single victimization survey cannot be used to measure change in crime because respondents tend to forget incidents that occurred in the more distant months and because they tend to misreport the date of events which they do remember. This error (called telescoping) is not randomly distributed throughout the recall period, but instead contains a bias so that the date given to the interviewer is more recent than when the crime actually occurred. Therefore, a single survey divided into monthly segments will always overestimate the true increase in the frequency of offenses.

Although victimization surveys are the best known type of survey conducted within the criminal justice system, there are several other purposes that could be served by well designed and executed survey research efforts. One such purpose would be to survey a sample of the general population in order to assess the relative importance of

alternative policy choices that could be made by the criminal justice system. Alternatively, the same survey could examine the relative importance of criminal justice goals when compared against other kinds of public policies. Surveys of this type could be used to ascertain the perceived seriousness of different kinds of crimes.

In *"Measurement Strategies for Determining Citizen Policy Preferences"* several of the more important issues and technical procedures are described. Although the discussion is generally confined to measuring citizen preferences about criminal justice goals vis a vis other policy areas, most of the techniques could be used in surveys conducted for different but related purposes.

TECHNIQUES FOR MAKING EVALUATION USEFUL

It was noted at the beginning of this introduction that evaluation has two primary purposes: (1) the production of scientifically valid information, and (2) the production of information that will be useful in planning and decision making. Each of the five papers in Section 3 of the Handbook is intended to present information or describe procedures that will improve the utility of evaluation for planning and decision making.

"An Introduction to Evaluation for Planners and Decision Makers" describes the role of evaluation in the planning process, identifies the different kinds of evaluation that could be conducted, and provides other information that should be useful to planners and decision makers in their efforts to insure that their informational needs will be met by the evaluation report.

The typology of evaluation described in the paper is the one currently being used by LEAA in its Evaluation Training Course. The different kinds of evaluation are identified by whether the final performance criterion is an outcome (a major social consequence), a result (an intermediate effect), or an activity (something done by the project itself).

Impact Assessment establishes the causal relationship between outcomes (such as crime reduction) and the activities or results of the project.

Process Evaluation establishes the causal relationship between

results (such as an increase in arrests) and project activities.

Monitoring examines the activities or activity levels of the project and relates these directly to the resources invested in the project.

The LEAA typology is extended so that it also incorporates an explicit identification of what the project is being compared with. Thus, the project as a whole could be compared against some other alternative strategy (a "black box" evaluation) or several project activities/components could be compared against each other (a project component evaluation). The position presented within the paper is that the type of evaluation that should be conducted depends upon the questions that need to be answered in order to meet the future informational needs of planners, project directors, and/or other decision makers. The questions that need to be answered depend on the decisions that will have to be made by the key audience of the evaluation findings.

Another section of the paper describes the role of the evaluator in project planning and the role of planners or project directors in the conduct of the evaluation. It might seem reasonable to suggest that it is the evaluator's task to produce scientifically valid information and it is the task of the planner, project director, or other decision makers to operate the project and decide what questions should be answered in the evaluation. The suggestion presented in the paper, however, is that these tasks are so interrelated that efforts to completely separate them generally will result either in invalid answers to important questions or valid answers to trivial questions. The alternative is to have the evaluator involved (along with planners, project

directors, and other decision makers) before the project is implemented in order to insure that reliable data will be collected, the appropriate design can be implemented, and the relevant questions can be answered.

The paper *"A Systems Approach to Evaluation"* presents a strategy for determining the questions and propositions which should be included in the evaluation.

The first step is to describe the project as an interrelated system consisting of inputs, activities, results, and outcomes. The second is to use that description to trace the logic or theory of the project in order to determine why it is reasonable to believe that the inputs will indeed produce the activities at the levels expected, why these activities can reasonably be expected to produce the desired results, and why the results can reasonably be expected to produce the desired outcomes. The critical assumptions and intervening variables are identified through this process.

In the third step the different types of evaluation that could be conducted are related to the kinds of comparisons that could be made. Four different dimensions of performance then could be applied: the quantity, quality, timeliness, or cost (of outcomes, results, or activities).

If fully developed and applied to the specific project described in the systems diagram, this procedure would identify virtually all of the questions that might be addressed in the evaluation. The evaluator (or person responsible for designing the evaluation) would then need to ascertain the costs of answering these questions, identify those

most likely to be important in future decision making, and, through a process of discussion and negotiation with other relevant actors, arrive at a final agreement about the propositions or questions to be included in the evaluation.

DETERMINING THE "SUCCESS" OF A PROJECT

One of the most frequently discussed issues in criminal justice planning and evaluation is what constitutes "success" for a project.

The first paper in this series, *"Alternative Approaches for Establishing the Criteria of Success,"* examines the issue from a practical point of view. The determination of "success" requires an identification of the problems on which the project was supposed to have an impact, selection of measures for those concepts, selection of a particular amount of the problem that must be solved, and selection of a specific probability level for ascertaining whether the apparent effects of the project were due to a chance occurrence. The topics covered in this paper include the options available to project personnel for stating their goals and objectives in quantitative terms and the options available to the evaluator for converting these statements into propositions that are testable. In addition, the discussion presents a non-technical review of how tests of statistical significance should be reported and used in evaluation research.

From a more philosophical perspective, one could argue that there are two fundamental ways of determining the "success" of any government action. The first of these, a "responsive government approach," is based on the rationale that a "successful" policy is the one that the citizens would choose to continue if permitted to vote on the relevant options in a fair election. The second, a cost-benefit approach, is based on the philosophy that a "successful" policy is one which provides at least one more unit of value to the public for a one unit

expenditure of resources. These two approaches are often discussed in conjunction with one another, so that the public, through elected representatives, is permitted to identify the broad goals of government action, but technical information of a cost-benefit nature is to be used to determine the means for achieving those goals.

It is recognized, of course, that the decision making process is not nearly as responsive as that required by the "responsive government" approach, nor is it as rational as that envisioned in the cost-benefit approach.

The paper "*Cost Benefit and Cost Effectiveness Evaluation*" describes the step by step procedures for conducting cost benefit and cost effectiveness analyses. Although a cost benefit evaluation is the best possible kind, it is virtually impossible to obtain the data and design needed to assess the costs or benefits of social service programs. Cost effectiveness evaluations, however, can be conducted in most situations. These compare the per unit cost for two or more alternative strategies in achieving specified results or outcomes. It should be emphasized that cost benefit and cost effectiveness techniques require that the causal linkage between the project and the results or outcomes be established. Therefore, these types of evaluations do not in any way reduce the need for strong evaluation designs, and reliable data.

"*The Role of Evaluation in Rational and Bargaining Decision Making Process*" contains a brief but interesting description of how evaluation findings are used within each of these types of decision making procedures.

APPLICATIONS OF PROBLEM SOLVING TECHNIQUES IN CRIMINAL JUSTICE EVALUATIONS

Section 4 of the Handbook contains eight evaluation reports or excerpts from evaluations, each of which was selected because it demonstrates the use of one or more problem solving techniques that would be of interest and value to other evaluators. Introductory comments which identify or expand upon the particular techniques of major interest have been prepared and precede each of the reports or excerpts. The discussion, at this point, of each evaluation report will be limited to a very brief overview of the techniques that were particularly interesting.

"The Hidden Camera Evaluation Report" illustrates a very useful method for achieving random assignment in field conditions. In contains a well developed cost effectiveness component and demonstrates the value of using different designs and different comparison groups to rule out virtually all of the alternative explanations for the observed effects of the project.

"The Bellevue Citizen Involvement in Burglary Prevention Evaluation Report" demonstrates many of the typical problems encountered in field research. One of the problems was how to simultaneously analyze data from the treatment and comparison areas and the second was how to determine when the project effects should be expected to occur (i.e., when did the project "strat"). The discussion of this paper includes a rather detailed presentation of how additional analysis of the time series data would illuminate the conclusions drawn in the evaluation and provide

additional information on the effects of the project.

"The De-Institutionalization of Status Offender Project Evaluation" demonstrates how multiple regression and time series analysis can be combined to test the impact of the project on recidivism rates of the youths. A problem encountered by the evaluators was in ascertaining when the project "started." The time series excerpt illustrates the use of a double intervention point: one to test for changes that occurred when the court approved, in principle, the application for the grant and the second to test for significant changes when the project was implemented.

"The Target Hardening Evaluation" contains an informative discussion of reliability and validity problems in the measurement of burglary rates and demonstrates how one can proceed to assess the amount of error in the data. As with several of the other evaluations, the use of multiple indicators of performance, several different designs, and several different comparison groups greatly strengthened the confidence in the conclusions that were drawn. The evaluation also demonstrates a type of design that can be used to test for crime displacement effects.

"The Seattle Community Accountability Program Evaluation" includes a particularly thorough linkage statement that not only describes the theory of the project, but uses the theory as a guide for the selection of performance measures. The assumptions and limitations of it. (The actuarial table procedures are no longer being used by the Seattle evaluators.)

"The Burglary Prevention Team Project," according to the evaluation

report, was neither designed nor operated in such a way that it could be evaluated. This report was included because it describes why the project could not be adequately evaluated. An interrupted time series design was used and served the purpose of demonstrating that a simple pre-post comparison of burglary rates would have produced an erroneous conclusion.

"The Burglary Task Force Evaluation Report" illustrates the relevance of a "project component" evaluation in which two strategies used by the same project are compared in terms of several performance measures. In addition, a multiple time series analysis is used in an effort to rule out alternative explanations for the apparent results of a single time series analysis. The report is especially well presented in that it contains a very short summary on the first page, a good description of the project theory and rationale, and concludes with a brief but informative discussion of the relevance of the findings for project operation.

"The Driving While Intoxicated Impact Grant Evaluation" demonstrates the use of interrupted time series and two other rather innovative techniques. One of these is the use of regression equations to assess change in the productivity levels of police officers and the other is a lagged (time series) regression analysis to examine the direction of a cause and effect relationship.

FOOTNOTES

1. See Donald T. Campbell and Julian C. Stanley, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH (Rand McNally & Company, 1966).

2. See the *"Hidden Camera Evaluation Report"* in Section 4.

SECTION 2

TECHNIQUES FOR OVERCOMING TECHNICAL PROBLEMS IN EVALUATION RESEARCH

Abstract

The papers in this section discuss procedures for overcoming technical problems in evaluation research. The issues of design, analysis, and measurement are given primary attention.

SECTION 2A

A REVIEW OF THREATS TO VALIDITY*

Abstract

A review of the major types of threats to validity (i.e., alternative explanations for findings) is presented in this paper, with specific applications to the types of problems encountered in criminal justice evaluation. Following a short introduction and discussion of the meaning of each threat to validity, there is a one-page summary for each threat containing a short definition, an example, alternative approaches for solving the problems (if any exist), and a diagram containing the types of designs relevant to the discussion.

*This paper is a revision and expansion of materials presented by Anne L. Schneider and L.A. Wilson II at a special forum for evaluators in the State of Washington.

A REVIEW OF THREATS TO VALIDITY

Policies, projects, treatments, and other similar actions are supposed to contribute to the solution of social problems. In order to determine whether a project is effective in solving or ameliorating a problem, an evaluation must establish a causal linkage between the policy and the outcome measure. Thus, evaluation must go far beyond simply noting that a change in the level of the "problem" has occurred and must attempt to determine how much (if any) of the observed change in the level of the "problem" can be attributed to the independent variable which is of interest to the decision makers.

If something other than the activities of the project might have contributed to a change in the level of the problem (or to the difference between what was observed and what would have been observed without the project), then these other conditions "threaten" the validity of any conclusion that is drawn. Thus, the phrase "threat to validity" refers in a general sense to anything about the evaluation design or procedures that threatens the accuracy of a conclusion concerning the causal relationship between the independent variable and the dependent variable.

In any particular evaluation situation there may be dozens of "threats" to the validity of a conclusion, but the more common types have been identified and described by Campbell and Stanley and extended by Cook and Campbell.[1]

A study is said to have internal validity if one can determine that the treatment was the causal agent and all other alternative explanations for the observed outcome are eliminated. External validity refers to the

generalizability of the final results. The external validity of most evaluations is extremely low because even if it can be determined beyond a reasonable doubt that the program had a positive impact (internal validity), there is often no way to know whether the same program would have the same results in another city, with a different type of client, with a different director, and so on. It is also true, however, that the first step in producing knowledge about program effectiveness is to focus on internal, rather than external, validity.

Threats to the internal validity of an evaluation, as outlined by Cook and Campbell, are applied to criminal justice evaluation in the discussion below.[2] It should be kept in mind that many of these "threats" pertain to the group or area that received the treatment. Some of them are ruled out with the use of a control group, whereas others are not ruled out except under conditions of random assignment.

History: "History" is a threat to internal validity when an observed effect might be due to some event which took place between the pre-test and post-test and when this event is not the treatment of research interest. In this context, pre-test refers to data collected about or from an individual or an area prior to the intervention of interest. Post-test refers to data collected about or from individuals or areas after the intervention. Thus, all events which occur at about the same time as the treatment of interest are potential threats to the internal validity of the study. In addition, an event could affect the pre-test observations but not the post-test, or could affect both but in different ways. This, too, will confound the interpretation of change from pre to post.

Maturation: Maturation refers to the fact that individuals who are being studied, or areas that are being studied, change naturally over time. Individuals grow older, wiser, stronger, and so on. Areas of a city also change over time, although the type of change may not be so easily described as changes of individuals. Maturation is a threat to internal validity because these "natural" changes could produce the observed effects and the effect could erroneously be attributed to the treatment.

Testing: Some evaluations involve the administration of tests to persons before and after the intervention. Testing is a threat to internal validity because individuals can "learn" from having taken the test the first time, or, for some reason, having taken the test once influences the scores the second time.

Instrumentation: A change in the measuring instrument that is used to collect pre and post intervention data about individuals or areas could produce changes in the results that would be incorrectly attributed to the treatment. This is a particularly critical problem in criminal justice evaluation because much of the data is obtained from records that are kept by persons other than the eavluator and are subject to changes either in form or in policies concerning how the records are kept.

Regression to the Mean: Regression to the mean (also called statistical regression) is a threat to internal validity when the treatment of interest is used only on persons or areas that are especially "high" or "low" on the phenomenon being studied. It is normally the case that persons or areas which are abnormally "high" or "low" will

return to a more normal condition, over time, with or without a "treatment" being administered. The reasons for this differ, depending on the type of data being used. If a program is designed to provide special services to areas of a city only when those areas suddenly have an increase in the crime rate, then a decline in the crime rate without the intervention could be anticipated simply because the sudden increase was "abnormal." (This does not mean that all decreases from suddenly high crime rates are due to regression to the mean; it simply means that the researcher always must examine this possibility.) Studies involving individuals who are tested and then placed into a treatment program if their scores are especially high or especially low must consider regression to the mean as a possible explanation for change because measurement error in the pre-test will result in some persons having higher scores than "normal" for them and some having "lower" than normal scores. The group with the highest scores would normally be expected to have a lower group average during a post-test, whereas the group with the lowest scores would normally be expected to have a higher group average in the post-test condition.

Selection: Selection is always a threat to validity unless the control and experimental groups are randomly chosen. Selection bias refers to the fact that differences in the types of persons in the treatment and comparison groups that existed before the treatment could produce the differences in results or the changes observed between pre-test and post-test.

Mortality: Mortality refers to the fact that certain types of persons may drop out of a particular treatment (or control group between

the pre-test and the post-test. Thus, the groups are composed of different persons at the post-test than at the pre-test. Differences observed could be due to who dropped out rather than to the treatment.

Diffusion or Imitation of the Treatment: When treatments involve informational programs and when the experimental and control groups can communicate with each other, the controls may learn the information and thereby may receive the treatment. The experiment thus becomes invalid because there is no treatment or control group in any functional sense, and the experimental-control difference at the end of the experiment will not reflect any real differences in the treatment experienced even if the treatment was very effective.

Compensatory Equalization of Treatment: When the experimental treatment provides goods generally believed to be desirable, there may emerge administrative and constituency reluctance to tolerate the focused inequality that results. Thus, other sources may provide funds or treatments to the (presumably) untreated group. Again, this results in a conclusion of "no effect" when there might have been one.

Local History: Local history refers to events that happen only to the treatment group or to the control group, but not to both. (This actually is not a different threat than that discussed under "history," but Cook and Campbell in their later work have begun to distinguish between local history which affects one or the other group--but not both--and "global" history which affects all persons or groups.[3]

In any particular policy area there are likely to be additional threats to validity not covered by any of the more commonly discussed ones--including those presented above. We have added one to the list

because of its common occurrence in criminal justice evaluations:

Multiple Effects of Treatment: Multiple effects of a treatment or intervention can be a problem when not all of these are measured and when one of the effects confounds the measurement or interpretation of the outcome of interest. One of the common multiple effects in criminal justice occurs when program simultaneously increase the reporting of crimes and decrease the occurrence of crimes. If the reporting rate is not measured (pre and post), the change in it totally confounds the interpretation of change in the reported frequency of the crime.

The major difference between true experiments and quasi-experiments has to do with internal validity and with the likelihood that one or more of the threats listed above will confound the results of the study. A true experiment, defined here as random assignment to treatments, avoids most of the threats. When respondents are randomly assigned to treatment groups, each group is similarly constituted (no selection or maturation bias); each experiences the same testing conditions and research instruments (no testing or instrumentation problems); there is no deliberate selection of high and low scorers on any tests except under conditions where respondents are first matched according to, say, pretest scores and are then randomly assigned to treatment conditions (no regression to the mean problems); each group experiences the same global patterns of history; and if there are treatment-related differences in who drops out of the experiment this is interpretable as a consequence of the treatment and is not due to selection. Nevertheless, experimental cases which are not available for the post-test could differ substantially from control cases which are not available for the post-test.

Thus, randomization takes care of most, but not all, of the threats to internal validity (remaining are diffusion or imitation of treatment, compensatory equalization of treatment, local history, multiple treatment effects, and--in some situations--mortality).

With quasi-experimental groups the situation is much different. Instead of relying on randomization to rule out most internal validity threats, the investigator has to make them all explicit and then rule them out one by one.

Making explicit and then ruling out the alternative explanations for the observed results is a difficult process. The investigator must think through each of the different threats and determine how each might have accounted for the observed results. When possible, alternative explanations should be tested empirically in the same way that the original effects of the treatment were tested. The best procedure is to anticipate the likely threats to validity and design the evaluation in such a way that data will be available to test the plausible alternative explanations.

On the subsequent pages, several of the more common threats to validity are defined, an example relevant to criminal justice is provided, and the extent to which the problem is "solved" by a randomly selected control group, a non-equivalent comparison group, and an interrupted time series design is examined briefly.

The following symbols are used in the discussion:

O_1 refers to a pre-project measurement on the dependent variable of interest, such as the crime rate of a city; scores on a test given to a group of persons, the number of offenses committed by a group during a specific time period, and so on

O_2 refers to a post-project measurement on the dependent variable of interest for the same group (or area) measured in the pre-project time period

X refers to the independent variable--the experimental treatment or project

(R) means that cases were randomly selected into the experimental group, which receives the treatment X, and into the control group, which does not receive treatment X

There are, of course, many other types of designs than those portrayed in the following tables. The purpose here is to clarify the threats to validity and to give the evaluator guidance as to how a selected number of designs deal with the problem. Designs that are referenced in the text or that are especially relevant to the discussion are portrayed on each page.[4]

HISTORY

Definition for Pre-Post Designs

An event other than the treatment could occur between the first and second measurements, or an event could alter the value of the pre-treatment observation (but not the post), or events could change both the pre and post observations but at different magnitudes of change.

Example for Pre-Post Designs

An event other than the project could occur between the measurement of the outcome at 0_1 and 0_2. If so, this event could be confused with the effect of the project.

An event of some type could alter the observation taken before the project starts (0_1) so that it is abnormally high or low. If this historical event does not continue with the same effect on the post-project observation 0_2, then 0_2 would have differed from 0_1 even if the project X had not been implemented.

Approaches to a Solution

Either of the experimental designs (cases are randomly assigned either to the experimental group which receives the treatment X or to a control group which does not receive the treatment) will suffice to rule out any historical threats to validity which have the same impact on both groups. An experimental design will not suffice to rule out events which affect one group but not the other after the random assignment.

A comparison group design (cases are not randomly selected) also will suffice to rule out any specific historical event that affected both groups in the same way but will not control for events that influenced one group but not the other.

In an interrupted time series design, many historical events occurring during the pre-project time period and similar events occurring close to the intervention X which are of a similar magnitude and have a similar type of effect could be ruled out. Specific events occurring at exactly the same time as X which never occurred before or which are of a different magnitude or which affect the observations in a different direction cannot be ruled out with this design. The multiple time series design will rule out any events that affect both the treatment group and the control group.

Pre-Post Design
0_1 X 0_2
Pre-Post Experimental Design
(R) 0_1 X 0_2
(R) 0_1 0_2
Post-Only Experimental Design
(R) X 0_2
(R) 0_2
Pre-Post Comparison Group
0_1 X 0_2
0_1 0_2
Single Time Series*
0000 X 0000
Multiple Time Series*
0000 X 0000
0000 0000

* One should have at least 10 pre-program time points for these designs and more if possible. One or more post observations are needed.

REGRESSION TO THE MEAN

Definition

Groups or areas that have extreme scores at one point in time tend to revert toward the average of the population from which they were drawn at subsequent points in time. Regression to the mean is a problem when clients or areas with extremely high or low values at O_1 are selected for treatment.

Example in Pre-Post Design

If the project selects only those cases with the highest scores on the pre-test (O_1) then these clients will tend to have lower scores on the post-test even if the treatment had not been given. If the project selects only cases with the lowest scores at O_1, the scores would be expected to increase by O_2 even without treatment.

Approaches to a Solution

Either type of experimental design will suffice. The evaluation could identify a group of persons (or areas) that "need" the treatment X and then randomly assign some to it and others to the control group. Even though the scores will change because of regression effects, the changes will occur for both groups and if there is a difference in the change or in the value of O_2 between the groups, it could be attributed to X.

Comparison designs generally are not sufficient to rule out regression to the mean if the treatment group takes most or all of the cases with extreme scores. One or more pre-program comparison groups or areas that did not receive treatment but had experienced equally extreme scores in the past and for which a second measure is available for about the same time lag as the treatment group could be used to estimate the regression effects. (This has been called a "different time" comparison group design.)

Time series designs with many pre-program observations control for regression to the mean if the groups or areas had experienced scores as high as those that occurred for the treatment group. This permits an estimate of the amount of regression to the mean which would be expected.

Pre-Post Design

O_1 X O_2

Pre-Post Experimental Design

(R) O_1 X O_2

(R) O_1 O_2

Post-Only Experimental Design

(R) X O_2

(R) O_2

Pre-Post Comparison Group

O_1 X O_2

O_1 O_2

Different Time Comparison

O_1 O_2

O_1 X O_2

Single Time Series *

0000 X 0000

Multiple Time Series *

0000 X 0000

0000 0000

*One should have at least 10 pre-program time points for these designs and more if possible. One or more post observations are needed.

MATURATION

Definition

Persons within the groups or areas that receive the treatment are getting older, more mature, wiser, more experienced, or changing in some other way through time.

Example in Pre-Post Design

Because there is a time lag between O_1 and O_2 and because the people who receive the treatment X are getting older or wiser or more experienced, the value of O_2 would be expected to change even without the intervention of the project X.

Approaches to a Solution

Random assignment of persons from an eligible group into the treatment and into the control group will control for maturation effects because, whatever these are, they influence both groups in the same way. Thus, if the change between O_1 and O_2 is greater for the experimental group than for the control, it could be attributed to X rather than to maturation. Or, if O_2 differs between the control and experimental groups this difference could be attributed to X.

A comparison group that is equivalent to the treatment group in terms of age or experience or other characteristics that change with time which might influence the value of O can be used to estimate the effect of maturation and to determine whether this threat has confounded the interpretation.

Time series designs with many pre-project observations on the same groups or areas which later receive treatment will control for maturation effects if these effects are linear through time; but time series will not control for non-linear maturation effects. Time series designs with many pre-project observations on different groups or areas from those which later enter treatment but which are about of the same age (and so on) will control for maturation effects. Multiple time series also controls for maturation if the comparison group or area is equivalent (age, experience level, etc.) to the treatment group or area.

Pre-Post Design

O_1 X O_2

Pre-Post Experimental Design

(R) O_1 X O_2

(R) O_1 O_2

Post-Only Experimental Design

(R) X O_2

(R) O_2

Pre-Post Comparison Group

O_1 X O_2

O_1 O_2

Single Time Series *

0000 X 0000

Multiple Time Series *

0000 X 0000

0000 0000

* One should have at least 10 pre-program time points for these designs and more if possible. One or more post observations are needed.

SELECTION

Definition

Criteria used to select persons into the treatment group may differ from the criteria used in selecting persons for the comparison group.

Example for Post-Only Comparison Group Design

In the post-only comparison group design, there is no pre-project measurement. If the criteria used to select persons for the treatment are such that persons entering could be expected to do "better" or "worse" than those in the comparison group, then the value of 0_2 would differ between the groups even if the treatment had not been given. Self-selection into treatment produces the same problem because those interested in receiving treatment probably differ from those not interested and the "not interested" group would constitute the control group.

Approaches to a Solution*

Random assignment from a group of eligible persons into the treatment and control groups will randomly distribute pre-project differences. The pre-post experimental design and the post-only experimental design will both solve the problem.

Pre-post comparison group designs do not solve the selection bias, although some analysis procedures are available that (in some conditions) will permit valid conclusions to be drawn. These procedures include multiple regression, actuarial tables, matched pairs, and multiple classification analysis.

Time series designs do not solve the selection bias problem but, again, there are procedures which can be used that (under some circumstances) will permit valid conclusions to be drawn.

Pre-Post Experimental Design

(R) 0_1 X 0_2

(R) 0_1 0_2

Post-Only Experimental Design

(R) X 0_2

(R) 0_2

Post-Only Comparison Group

X 0_2

0_2

Single Time Series**

0000 X 0000

Multiple Time Series**

0000 X 0000

0000 0000

---

* Two other papers in this handbook discuss the value and use of comparsion group designs, interrupted time series, and various analysis procedures in resolving the problem of selection bias. See "Comparison Group Designs," "Prediction Models," and "An Instroduction to Interrupted Time Series."

** One should have at least 10 pre-program time points for these designs and more if possible. One or more post observations are needed.

TESTING EFFECTS

Definition

Taking a test can have an influence on the scores obtained the second time the test is taken.

Example in Pre-Post Design

If taking the test the first time has an influence on the scores received the second time, then one would expect the value of O_2 to differ from O_1 even if the treatment had not been given.

Approaches to Solutions

Either of the experimental designs resolves the problem. In the pre-post experimental design one can assume that whatever testing effects exist will influence both groups in the same way. In the post-only experimental design there are no testing effects to worry about.

The pre-post comparison group design sometimes will solve the problem because whatever effects the taking of the test has on subsequent scores should exist for the comparison group if the latter is equivalent to the treatment group in age, intelligence, and so on.

In the time series designs where the same group is tested repeatedly during the pre-project time period, one would expect the testing effects to be captured in the pre-program trend and, since the expected post scores are projected from the trend, the design controls the problem. (If the testing effects are not linear, however, a linear projection of the trend will not solve the problem.) Time series utilizing a series of different, naturally occurring pre-program groups controls the problem if these historical groups are equivalent to the post-project groups. Multiple time series also control for testing effects if the comparison group is equivalent to the treatment group in terms of relevant variables such as those which influence the rate of learning.

Pre-Post Design

O_1 X O_2

Pre-Post Experinental Design

(R) O_1 X O_2

(R) O_1 O_2

Post-Only Experimental Design

(R) X O_2

(R) O_2

Pre-Post Comparison Group

O_1 X O_2

O_1 O_2

Single Time Series *

0000 X 0000

Multiple Time Series *

0000 X 0000

0000 0000

* One should have at least 10 pre-program time points for these designs and more if possible. One or more post observations are needed.

MORTALITY

Definition

Mortality is a biased and differential loss of cases from the treatment and control (or comparison) groups.

Example in a Pre-Post Experimental Design

Pre-Post Experimental Design			
(R)	O_1	X	O_2
(R)	O_1		O_2

Some of the clients randomly assigned to the treatment group will drop out of the program, producing a problem of whether to include them in the post-project measurement. Since no one "dropped out" of the "untreated" control group, the two groups are no longer equivalent. Or, if persons did drop out of the control group, these may not be the same types of people as those who dropped out of the comparison group. Another problem is introduced if clients are not available for the second observation. Since those not available in the experimental group may be different than the dropouts from the control group, the two groups are no longer equivalent.

Approaches to a Solution

There is no design that will solve the problem, not even the experimental designs. If the mortality problem is introduced because clients failed to complete treatment but were still available for post-test, then the evaluator should simply include the dropouts with the group to which they originally belonged. If the problem is because the O_2 measurement cannot be obtained, then the evaluator should attempt to insure that the dropouts from the treatment and control groups are equivalent to each other. (Some would recommend that each dropout be matched to a person in the other group and both excluded from the entire analysis.)

FOOTNOTES

1. Donald T. Campbell and Julian C. Stanley, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH (Rand McNally & Co., 1966), and Thomas D. Cook and Donald T. Campbell, "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings," in HANDBOOK OF INDUSTRIAL AND ORGANIZATION RESEARCH (Rand McNally, 1975).

2. Cook and Campbell (ibid.) added three "threats" to the list, including diffusion of treatment, compensatory equalization of treatment, and local history.

3. In subsequent parts of the paper local history is treated as a special case of the history threat.

4. In the time series designs, the 0000 actually refers to many observations (at least 10 in the pre-project period). Unless otherwise noted, it is assumed in the time series designs that the observations are on the same group or area. Distinctions are made among five different interrupted time series designs in the Handbook paper *"An Introduction to Interrupted Time Series Designs."*

SECTION 2B

COMPARISON GROUP DESIGNS*

Abstract

The logic of experimental and quasi-experimental designs is presented as a point of departure for examining the strengths and weaknesses of comparison group designs. Matched pairs, actuarial tables, and multiple regression constitute alternative procedures that an evaluator could use either to form a comparison group that is equivalent to the treatment group (matched pairs) or to statistically control for differences between the groups. The strength of a comparison group design depends on how similar the group is to the treatment group on whatever variables are relevant for ruling out one or more threats to validity. The general thrust of the paper is that comparison group designs, while not as rigorous as experimental ones, are better than pre-post designs, and in situations where the eligibility rules for entry into the project are quantitative and precise the comparison group design can be rather strong.

* Most of these materials are an expansion and revision of those presented by Anne L. Schneider at a special forum for Criminal Justice Evaluators in the State of Washington.

COMPARISON GROUP DESIGNS

Introduction

The most common typology of evaluation designs, popularized in Campbell and Stanley, includes experimental, quasi-experimental, and pre-experimental designs.[1] For our purposes, these will be defined in the following way:

1. Experimental designs are those in which persons (or areas) are randomly assigned to the various categories (or values) of the independent variable of interest. Although the phrase experimental design often refers to random assignment of cases into a treatment and non-treatment group the design is experimental if there is random assignment to two or more different types of treatment, two or more different types of counsellors (who might be using the same "treatment"), to several different amounts of a treatment, such as different number of hours in counseling. Randon assignment is the key to identifying an experimental design, not the "treatment vs. non-treatment" characteristic.

2. Quasi-experimental designs refer to a situation in which the groups (areas, or individual cases) differ in terms of their values on the independent variable of interest. Evaluators often have a pre-project or concurrent "untreated" group which could be compared to the project group. Less commonly found in the literature are situations where the evaluator has different amounts of a particular treatment (but no completely "untreated" cases). With this situation, the evaluator could, other things being equal, compare those that receive more with those that receive less in an effort to ascertain the optimal amount of a particular treatment.

Quasi-experimental time series, also called interrupted time series, refers to a design where the evaluator has several measures (usually 10 or more) before the treatment began and several observations after the treatment. The observations for the pre-project time period could be for the same persons (or groups or areas) that later enter the project; or they could be for different persons who form historical groups that would have been in the project if it had existed in the past.

3. Pre-experimental designs are those in which the evaluator has only the post-treatment observation on the project group or area (and nothing at all to compare these observations to) or has one pre-treatment and one post-treatment measure for the group that receives the treatment. One might notice that the commonly used pre-post design becomes an interrupted time series design when enough pre-treatment observations (about 10 or more) have been taken to establish a trend.

Comparison group designs are a type of quasi-experimental design, as defined here. More specifically, a comparison group design will be defined as one in which (at a minimum) the evaluator has post-treatment observations (and other relevant data, on a case-by-case basis) for the project group (or area) and has measures for another group on the same variables. Under this general rubric, then, there are several different types of designs: pre-post data on treatment and comparison group; post-treatment data on the treatment group and observations taken at the same time on a concurrent comparison group; or an historical comparison group with observations taken at one point in time but prior to the project implementation. The value of a comparison group depends on how equivalent (i.e., how similar) the group is to the treatment group and on whether one is juxtaposing this design against an experimental one, or a

pre-post, or some other type of quasi-experimental design. In general, the comparison group design is not as rigorous as the experimental, is much better than the pre-post, and is weaker than a multiple interrupted time series design. It should be noted, however, that the comparison group design can be extended by collecting case-by-case data at equally spaced time intervals on several pre-project historical groups and thereby converted into a type of interrupted time series design called a "different group" time series.

The Logic of Designs

Before discussing the problems with comparison group designs, it would be useful to review the logic of experimental (random assignment) designs, using a specific criminal justice example. Suppose that the purpose of a project is to reduce recidivism of juveniles, as measured by subsequent court contacts. At the juvenile court the youth will be assigned randomly either to the usual court procedure (C) or to the new program (T). At this point in time (i.e., when the youth enters the court system) there is some true probability that the youth later will recidivate. Every youth has a pre-treatment or prior probability of recidivating, but the evaluator does not know what that probability is. When the youths are randomly assigned to C and T the pre-treatment probability of recidivating is being randomly distributed between the two groups and, within sampling error, we know that each group's average probability of recidivating will be the same. In other words, before the treatment begins, we can be confident that the future expected recidivism rate of C (control group) is approximately the same as that of T (the treatment group), within sampling error.

When random assignment is used the future (expected) probability of recidivating E(Y) depends upon the pre-treatment probability (Z), and the

treatment effect.

Thus, for T: $E(Y_T) = Z_T + \text{treatment}_T$

and for C: $E(Y_C) = Z_C + \text{treatment}_C$

And, because of the random assignment, we know that the prior probabilities Z_T and Z_C are the same, except for sampling error. Thus, the Z term drops out of the equation and we can anticipate that any differences we later observe on Y_T and Y_C (the six month recidivism rates, for example, of the two groups) are due only to the differences in treatment or to sampling error.[2] When the actual measurement of Y_T and Y_C are taken, it is the case that the control group's recidivism rate is being used as the estimate of what the treatment group's recidivism rate would have been if the treatment had not been received.

The fundamental principle in random assignment and experimental designs is that the scores of the randomly selected control group on the dependent variable can be used as the estimate of what the experimental group scores would have been IF THEY HAD BEEN IN THE CONTROL GROUP INSTEAD OF THE EXPERIMENTAL GROUP. Thus, the comparison between control and experimental scores provides an estimate of the effect of the treatment. The sampling error can be estimated from the size of the sample and if the differences between recidivism rates of T and C are greater than what would have been produced by error, one can conclude that the difference in scores is attributable to the treatment rather than to error.

This situation should be contrasted with what exists when we do not have random assignment. At intake, there is a prior probability for each youth that he or she will recidivate. Again, the evaluator does not know what the probability is. The intake officer, in this situation,

does not randomly assign youths to the project and to the traditional treatment, but instead uses his or her judgment as to which program would be best for the youth. The result of this, for the evaluator, is a complete lack of knowledge concerning not only what the prior probability of recidivism is, but also a lack of knowledge concerning whether the pre-treatment recidivism probability (Z_T, treatment) is at all similar to the pre-treatment probability (Z_C, the comparison group).

When the evaluator later measures the recidivism rates of T and C, he or she cannot simply compare the two and draw any conclusions about the effects of the treatment because the pre-treatment probability of recidivism cannot be assumed to have been equivalent across the groups.

General Approaches to Solving the Problem

It should be noted at the outset that there are no solutions to the problem of drawing conclusions about treatment effects when cases were not assigned randomly to the treatment and comparison groups.

Matching is one of the commonly used techniques for solving the problem of non-random assignment. In some types of research, the evaluator may have sufficient control over the situation so that whenever the project selects a case for treatment, the evaluator would have an available pool of eligibles from which he or she could select a matched case for the comparison group. This situation is more typical, perhaps, in some types of educational programs or clinical psychology, than in criminal justice. The procedure is for the evaluator to select a case from the eligible pool that "matches" the one selected for treatment in terms of whatever variables are presumed to affect the dependent variable of interest. Thus, in education

if a pre-test has been given, the evaluator normally would select for the comparison group a case that matches the one in the treatment group on the pre-test score and perhaps on some other variables that presumably would influence learning (if that is the dependent variable) such as age, race, sex, family stability, and so on. In criminal justice, the evaluator rarely has the chance to randomly select from the pool of eligibles one of the cases that "matches" the one placed in the program. Instead, the matching procedure generally is done post-hoc. From the group of persons not selected into the project, the evaluator attempts to match on whatever characteristics he or she thinks the project used in determining who would receive the treatment and on any other variables that might influence the dependent variable of interest.

The evaluator has two procedures for "matching" cases in order to create a comparison group. One of these, called matched pairs, requires the evaluator to randomly select one case from the non-project group that matches each case in the project on variables thought to be important. For example, suppose the project accepts a client who is a male first offender charged with burglary who lives with his stepfather and has been expelled from school. The evaluator, using matched pairs, might decide that all of these other characteristics are important and identify (in the non-project group) all 16 year old first offender males, charged with burglary, living with a stepfather, who have been expelled from school and then select (randomly) one of these cases for the comparison group.

The second procedure that could be used by the evaluator is to select all (or many) non-project cases that meet these criteria and use the proportion of this subset who recidivate as the estimated probability of recidivating for all project clients who fit into that same category. This tech-

nique yields an actuarial table or set of tables from which the expected recidivism of the treatment group can be estimated. It clearly is superior to the matched pair procedure because it would provide a much more stable estimate of the probability of recidivating than would a single case. Even so, most researchers would argue that a multiple regression approach would be superior to the second matching procedure unless the number of cases within each subset (e.g., 16 year old male, first offender, charged with burglary) is relatively high.[3] A discussion of why multiple regression normally would be superior is beyond the scope of this paper, but in general it has to do with the fact that multiple regression analysis provides the best estimate of the value of the dependent variable (recidivism, for example) for each independent variable, holding each of the others constant. The actuarial-table approach permits any type of interaction among the independent variables to alter the recidivism probabilities and permits any type of non-linear relationship that exists in the non-project group to be used as estimates even when the interaction and/or the non-linear relationship is produced entirely by error variance attributable to the small number of cases in the cells. (Multiple regression analysis can incorporate non-linear relationships and interaction terms can be included in the equation. Whether these are useful in terms of predictive accuracy would be determined with tests of significance on the regression coefficients of the non-linear terms and the interactions terms.)*

---

* There are special types of regression analysis which could be used and which, in some circumstances, would be superior to multiple regression analysis. One of these, multiple classification analysis (MCA), is similar to the actuarial table approach except that tests of significance are made for the cell estimates and the procedure identifies the combination of independent variables that will yield the most accurate predictions

The extent to which any of these three approaches (matched pairs, actuarial tables, or multiple regression) will permit the evaluator to draw valid conclusions about project effectiveness depends on the actual procedures that were used to form the treatment and comparison groups. If cases were selected so that prior to the treatment, the project and comparison groups have widely divergent expected values on the dependent (outcome) measure, then none of the procedures will work very well. Nevertheless, a comparison group design will still be better than a pre-post design for reasons that will be explained below.

Suppose that the project conducts a pre-test of some type and then accepts the clients who have either the highest or lowest scores on that test (i.e., the "easy" or "hard" clients). The evaluator could select matched pairs based on the pre-test or could select matched pairs based on the test and other characteristics. Alternatively, he or she could use an actuarial table approach or multiple regression. Regardless of the technique, there almost certainly will be a serious regression-to-the-mean problem within the treatment group that does not exist for the comparison group. Clients whose scores were extreme in the pre-test tend to score closer to the average of the original population on a post-test even without any type of treatment. Thus, if the project takes the hard cases they will tend to do better on the post-test whereas if the comparison group consisted of easier cases they would not regress as much toward the mean on the post-test. Conversely, if the project selected the "easy" cases, they

(based on least squares estimates). Logit analysis is similar except that the researcher specifies the combination of cells and uses a logit transformation on the percentages within the cell in order to correct for ceiling effects.

would tend naturally to score worse on the post-test whereas if the comparison group had harder cases they would tend to do better on the post-test.

Similar types of problems occur if the persons responsible for selecting clients into the project use judgmental criteria that are related to the dependent variable of interest. Thus, if the intake officer in the first example selected persons for the project who were thought to have a high probability of recidivating (based on unmeasured characteristics, such as attitude) and if his or her judgment is at all accurate, then the treatment group would not be equivalent enough to the comparison group to draw conclusions even though the two were perfectly matched on characteristics such as age, race, and sex. Conversely, if the selection procedure were such that the "easy" cases were in the project and the harder ones in the control group, the comparisons of recidivism between the groups would not be valid.

If the evaluator knows or suspects that the project and comparison group differ in that the selection procedure tended to place "easier" cases in one and "harder" cases in the other, then the best procedure to use would be multiple regression but even this technique will not completely adjust for pre-program differences in the probability of recidivating.[4] Nevertheless, if the multiple regression procedure indicates that the project received the harder cases, and if this matches the information that the evaluator has concerning the selection procedure, and if the project recidivism rate is less than that of the comparison group anyway, (with the other confounding variables controlled in the regression equation) then the evaluator could conclude that the project is effective and the magnitude of its effect probably is underestimated. (This general-

ization is true only for a comparison of the post-test scores. If pre-post change scores are used, the evaluator must be confident that regression-to-the-mean problems are affecting both the project and comparison groups in the same way.) On the other hand, if the project is not shown to have a significant effect on the dependent variable (with the potentially confounding factors controlled in the equation) and if the project did receive the "harder" cases, then the evaluator cannot draw any conclusion at all about the true effectiveness.

The situation which has been described above is one in which the criteria for selecting cases into the project either are not known, or cannot be measured quantitatively, or are presumed to be very biased in terms of the expected outcome measure. Even in this situation, a comparison group design is better than a pre-post design because with the latter anything that affected either the first or the second observations (other than the treatment) could account for differences and there is no way to rule out any of them. For example, post-treatment scores can be expected to differ from pre-project scores because of such threats to validity as maturation (clients get older, wiser, etc.); historical events that affected either the pre or post-test; regression to the mean for persons with high or low scores on the pre-test; changes in the methods of collecting data; changes in policy decisions concerning how the variables are to be measured; or testing effects. A comparison group does not have to be <u>perfectly</u> equivalent to the project group in order to rule out some of these types of threates to the validity of the conclusions; it must be equivalent on the <u>relevant</u> variables. For example, a comparison group of juvenile delinquents who are of the same approximate age of delinquent youths in the treatment group could sometimes

be used to estimate the amount of change attributable to maturation. A comparison group composed of all delinquent youths except those in the project could be used to examine whether policy changes at the juvenile court or police department affecting all delinquents could account for pre-post changes in the treatment group. The application of common sense, creativity, and ingenuity in the use of comparison groups--even when the group is not perfectly equivalent--can help a great deal in ruling out alternative explanations for observed changes in the treatment group.

A second situation is one in which the project meets the following conditions:

1. It has precise, quantitative criteria that are used to select persons for treatment and follows these explicitly;

2. It accepts all cases that meet those criteria;

3. There is a relatively large pre-project group of persons for whom the data are available to determine whether they would or would not have been eligible for the project if it had existed at that time; and

4. There is a relatively large pre-project group of persons who would have been eligible.

If these conditions are met, the evaluator could use the matched pair procedure (selecting one eligible case from the pre-project group, randomly, for each eligible case in the project). Or, better, the evaluator could use all of the pre-project eligibles as a basis of comparison utilizing multiple regression analysis to control for any differences that exist between the groups. The groups, of course, would be identical in terms of whatever characteristics were used to define eligibility, but might differ on other characteristics because of shifts, over time, in the characteristics

of persons eligible for the project. One possibly confounding factor is that there have been gradual changes, over time, in the dependent variable (such as recidivism rates) so that differences in means between the pre and post groups represent only a continuation of the trend rather than an actual change attributable to the treatment. If so, the evaluator would be wise to select the pre-program group at equally spaced time intervals prior to project implementation and expand the comparison group design into an interrupted time series design (ITS). The ITS, in this instance, simply consists of a series of comparison groups each representing a particular time interval prior to the project. The scores on the dependent variable can be tested in the pre-project groups to determine if there is a gradual trend in them. If not, the entire pre series can be compared with the post series. A second confounding factor is that some historical event influenced the pre or post measures. The eavluator should examine the situation carefully to determine whether it is plausible to suspect that this happened. In spite of these potential problems, the use of a pre-program comparison group (or groups), matched perfectly in terms of eligibility criteria, using multiple regression to control for any other differences, constitutes a rather strong design.

One of the conditions mentioned above is that the project must take all of the eligible clients. If they do not, then the evaluator must ascertain the basis for their accepting some and rejecting others. If any of the criteria are judgmental and cannot be replicated in the pre-project comparison group, then the evaluator could compare the entire pre-project group of eligibles with the entire post-project eligibles even though some of the latter were not in the treatment group. This constitutes a rather

severe test of project effectiveness since the project did not take all the clients. In addition, the "treated vs. untreated" eligibles could be compared to determine which were the "harder" and "easier" cases. If this can be ascertained, then it may be possible to obtain useful information about project effectiveness in spite of the bias in the groups. As noted previously, if the project takes the harder cases and still does significantly better than the comparison group, the evaluator has a basis for saying that the project is effective. And, if the project takes the easier cases but does no better than the comparison group, there is no evidence that the project is effective. With other results, however, no conclusions can be drawn.

Still a third situation is that the project cannot take all of the eligible cases but selects on the basis of "first come, first served". This presents fewer problems for the evaluator than judgmental selection because it is likely that those who came too late to be accepted (or missed being in the project because it was full during the time they were eligible for it) did not differ much, if at all, from the persons accepted.

Summary

The major points in this paper can be summarized as follows:

1. Comparison group designs are not as rigorous as experimental designs, but are better than pre-post. They are generally weaker than interrupted time series but can easily be expanded into a combination comparison group, time series, design.

2. Matching, using the matched pair procedure, is not as good a technique for dealing with the problems of non-random assignment as is the use

of actuarial tables and the latter is not as useful, under most conditions, as multiple regression analysis.

3. In situations where the project has explicit, quantitative eligibility rules which can be replicated in a pre-project group, or if there is a group of eligibles whose only reason for exclusion from the project is that the project was filled to capacity when they were eligible for it, then the evaluator should use these cases (eligible but not in the project) for the comparison group. The evaluator should examine the groups carefully for any differences between them and, if one is a pre-project group, should also assess whether trends or historical events could constitute alternative explanations for any differences that might be observed.

4. In situations where the "easy" cases were assigned to one group and the "hard" cases to the other, the evaluator has to be extraordinarily cautious about drawing conclusions and none of the procedures (matched pairs, actuarial tables, multiple regression) will work very well.

FOOTNOTES

1. See Donald T. Campbell and Julian C. Stanley, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH (Rand McNally & Co., 1966).

2. This can be shown as follows:

$$E(Y_T) = Z_T + \text{treatment}_T$$

$$E(Y_C) = Z_C + \text{treatment}_C$$

and: $E(Y_T) - E(Y_C) = (Z_T + \text{treatment}_T) - (Z_C + \text{treatment}_C)$

since: $Z_T = Z_C$ the equation becomes

$$E(Y_T) - E(Y_C) = \text{treatment}_T - \text{treatment}_C$$

3. See Donald T. Campbell and Robert F. Boruch, "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects," in C.A. Bennett and A. Lumsdaine (eds.), CENTRAL ISSUES IN SOCIAL PROGRAM EVALUATION (Academic Press, 1975).

4. Re-analysis of Head Start program data, combined with the use of hypothetical data with known properties, have revealed that the multiple regression approach did not adequately adjust for pre-Head Start differences among youngsters. Thus, the conclusion that Head Start was not effective or, in fact, even harmful, is almost certainly erroneous. See Donald T. Campbell and A. Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations can Mistakenly Make Compensatory Education Look Harmful," in J. Hellmuth (ed.), COMPENSATORY EDUCATION: A NATIONAL DEBATE (Brunner/Mazel, 1970).

SECTION 2C

INTRODUCTION TO INTERRUPTED TIME SERIES DESIGNS*

Abstract

In interrupted time series the pre-project observations are used as the basis for estimating the trend and level of the post-project observations. Differences between the predicted values in the post period and those actually observed can be tested to ascertain whether a statistically significant change in the trend or level occurred at the time of the intervention. Five different types of interrupted time series designs are identified in the paper. The extent to which alternative explanations for the apparent impact of the project are controlled by these depends on which of the designs has been used and the nature of the threat to validity. The analysis of interrupted time series data presents eavluators with especially complex problems. This paper describes several statistical tests and procedures which can he used and explains how the evaluator should test for autocorrelation in the residuals. Another paper in the Handbook, "Applications of ARIMA and ANCOVA to Interrupted Time Series," contains a more technical discussion of these two fundamentally different approaches to the analysis of time series data.

*These materials are a revision and expansion of those originally prepared by Anne L. Schneider and L.A. Wilson II for a Special Forum of the Association of Law and Justice Evaluators in the State of Washington.

INTRODUCTION TO INTERRUPTED TIME SERIES DESIGNS

The purpose of this paper is to describe interrupted time series designs (ITS) and discuss how they can be used in criminal justice evaluations. The presentation is divided into three major parts: (1) A description of the underlying logic of interrupted time series designs, (2) a discussion of threates to validity in the quasi-experimental time series approach, and (3) a presentation of statistics for estimating the significance of various types of changes that might occur in the data.

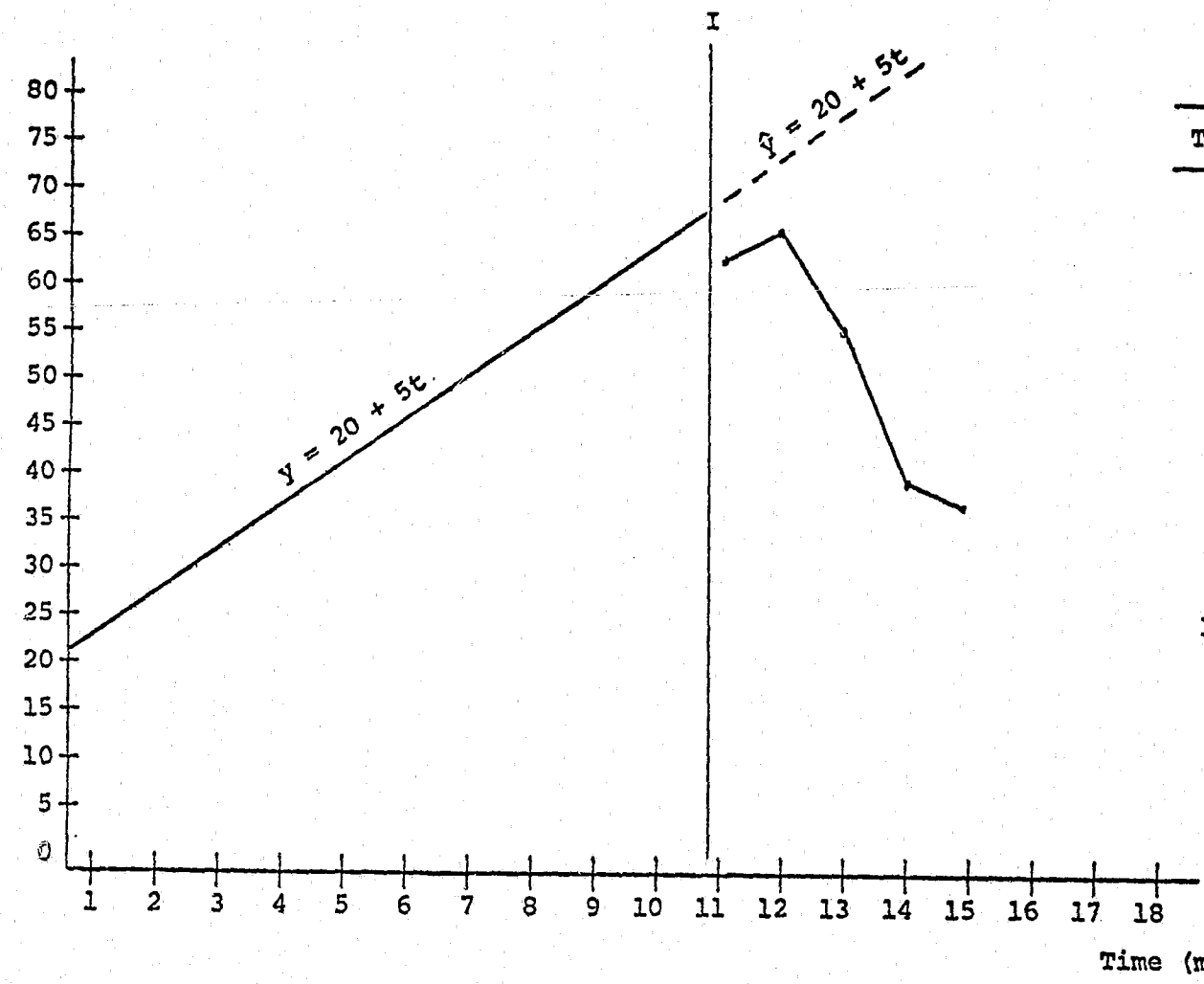
LOGIC OF INTERRUPTED TIME SERIES

Time series analysis, as that phrase is normally used by statisticians, economists, and others, refers to an analysis of a single variable measured at many successive time points. Interrupted time series analysis refers to an analysis of a single variable measured at many successive time points, but with some of the measures taken prior to the intervention of a program or policy and other observations taken after that intervention. Figure 1 shows a set of data that could be used in an interrupted time series analysis.

On the horizontal axis are time points (months) with "one" referring to the most distant month for which a measure of armed robberies has

FIGURE 1

INTERRUPTED TIME-SERIES



D A T A	
Time	Y
1	25
2	30
3	35
4	40
5	45
6	50
7	55
8	60
9	65
10	70

been obtained and "15" referring to the most recent month for which there is a measurement. On the vertical axis, in the example, is the number of armed robberies that occurred for each of the months. The vertical line marked "I" refers to the intervention of some type of policy at the end of the tenth month.

There are three ingredients for any interrupted time series design:

1. A minimum of 10 observations of the dependent variable prior to the intervention, and at least one observation afterward. At least 10 observations are needed to obtain a stable estimate of the trend in the data.

2. The observations must be taken at different times. The time unit can be in days, months, quarters, or years, but it generally is better to narrow the time interval in order to increase the number of time points rather than to aggregate to a larger interval (years, for example) and reduce the number of time points.

3. The third essential ingredient is that one must know when the intervention took place. If information is not available about when the program was implemented, then it is very difficult--sometimes impossible--to draw inferences from the analysis.

The basic logic of interrupted time series is rather straightforward. Given that the dependent variable (Y) has been measured before and after the intervention, and that the observed values of Y have been obtained (shown in Figure 1 as the solid line), then the key question is: What would Y have been after the intervention if the intervention had not occurred? Almost all research seeks to compare one or more observed outcomes with some theoretical expectation of what the outcome

would have been if X (the intervention) had not occurred.

In interrupted time series analysis the expected values of the outcome (the dependent variable) after the intervention are obtained by projecting (forecasting) these values from the pre-program data.

As shown in Figure 1, the data before the intervention follow a perfect linear upward trend, increasing by about five robberies per month during the 10 months for which there are measures. The dotted line after the intervention represents the expected scores on Y if the intervention had not occurred. The solid line after the intervention represents the observed values of Y.

How are the data projected to obtain the expected values?

Linear projections, using regression analysis, are acceptable for most of the shorter time series.[1] Time, measured by integers (one, two, three...), is used for the X variable in a normal least square regression equation.

There are three patterns of change that should be watched for and tested for in interrupted time series analysis:

1. Long-term change in the trend (which is measured as the slope of the line--that is, the regression coefficient);

2. Short-term change (which is indicated by a shift in the level of the series right after the intervention point); and

3. The durability of the change (which is determined by the slope of the post-intervention time points).

Several patterns of change are shown in Figure 2. Figure 2a indicates that the program had an immediate impact and also reversed the pre-program trend from an upward trend to a downward trend.

FIGURE 2

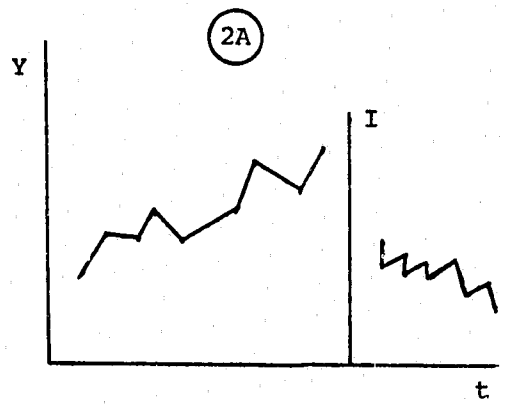
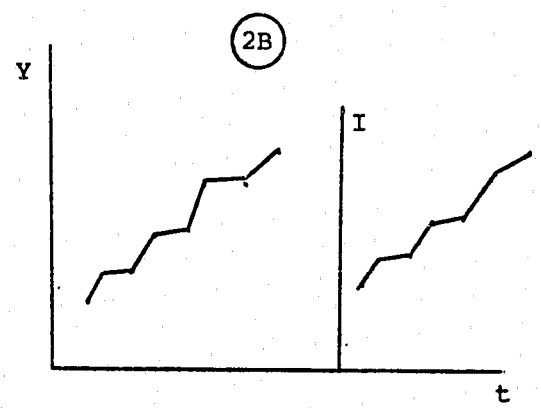
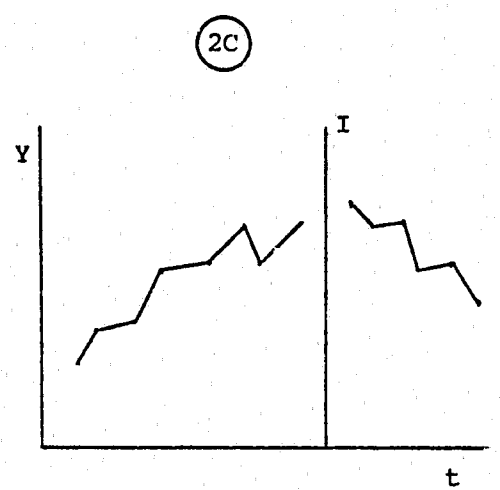
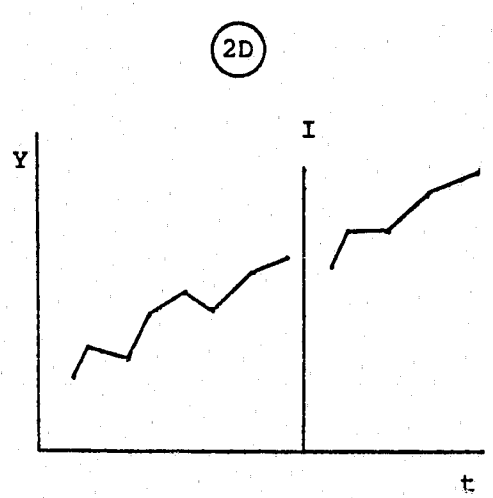
	LONG TERM CHANGE OCCURRED	NO LONG TERM CHANGE OCCURRED
SHORT TERM CHANGE OCCURRED		
NO SHORT TERM CHANGE OCCURRED		

Figure 2b shows that the program had an immediate effect in reducing the level of the series, but it did not alter the trend. This is sometimes called a step function.

Figure 2c shows that the program had no immediate impact on the level, but apparently altered the trend in the data. This type of pattern is especially difficult to interpret and to attribute to the program itself. It is wise to extend the pre-program data back for as many additional months as possible in order to determine whether this type of change has occurred in the past even when there was no intervention.

Figure 2d shows no change at all.

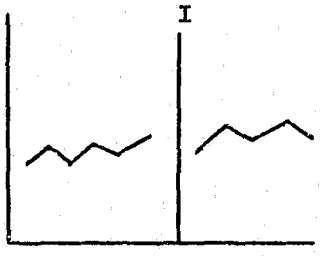
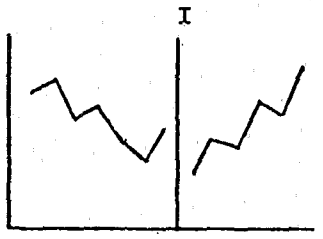
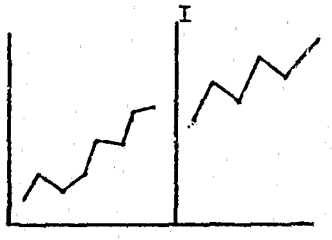
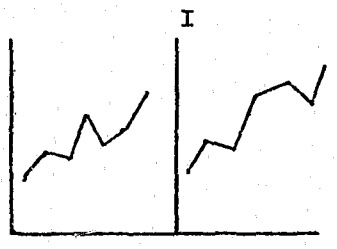
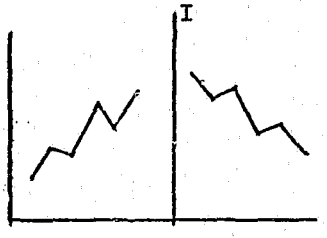
The basic logic of time series analysis is similar to that used in many other types of research designs. It is especially instructive to show the similarities between interrupted time series and pre/post designs with no control groups.

Suppose we have one year of pre-program data on the number of status offenders in the community who were detained at the juvenile court and one year of data after a status offender diversion program became operative. A pre/post design, with no control group, would take the average number of status offenders detained per month prior to the program and compare with the average number detained per month afterward.

An interrupted time series design would break the data out on a month-by-month basis and plot it for each month.

Under what conditions will the results of a pre/post comparison be the same as the results of an interrupted time series? Under what conditions will one make unwarranted inferences (or no inferences at all) with the pre/post design that would not be made with the time series design? Several conditions are shown in Figure 3.

FIGURE 3

	conclusion if using pre/post design	conclusion if using time series design
 (1)	n.s.	n.s.
 (2)	n.s.	significant
 (3)	significant	n.s.
 (4)	n.s.	significant
 (5)	n.s.	significant

Situation 1: In situation 1 (Figure 3) there is no trend either in the pre or post-program data. In this situation, one would draw the same conclusion regardless of the design that is used. This is, however, about the only situation where the same conclusion would be drawn.

Situation 2: There is a downward trend in pre-program months, followed by an immediate drop and then an upward trend in the post-program time period. The means (or totals) of pre and post would be almost identical. The pre/post design will obscure what really happened.

Situation 3: This diagram shows a steady and rather rapid upward trend. The pre/post design will show a significant effect because the means would differ substantially, when in fact there is no justification for it.

Situation 4: The pattern shown in situation 4 is a step change. The pre/post design will show no effect when, in fact, the entire level of the series is lower than it would have been without the program.

Situation 5: A change in trend without an immediate rise or lowering of the series after the intervention is shown in situation 5. Again, the comparison of means using a pre/post design will miss the relevant change which occurred.

Interrupted time series designs are not limited, however, to situations in which there is no control or comparison group. Several varieties of interrupted time series designs are described below.

1. Multiple Interrupted Time Series: In this design the evaluator includes a comparison group for whom a series of pre and post measurements on the dependent variable are taken at the same time intervals as used for the treatment group or area.

2. Experimental Interrupted Time Series: In this design cases

or areas would be randomly assigned either to treatment or non-treatment conditions and measurements would need to be available for both during an historical pre-project time period and after the project was implemented.

3. Different Group Interrupted Time Series: In most references to interrupted time series designs, it is assumed that the pre-project measures have been taken for the same group or area that later enters treatment. Examples would include time series analysis of the crime rate for a city or area within a city, the historical offense rate of offenders who later enter the project, and so on. For some types of projects, however, there will be naturally occurring groups of persons who entered the system prior to the treatment but exited before the project began. In these situations, the naturally occurring groups could be divided in accordance with when they entered the system and composed into a series of historical groups. Juvenile offenders who had contact with the juvenile court in January 1974, for example, could become the first set of observations; those entering in February 1974 would be the second; and so on for each month prior to when the project was implemented. The proportion of youths detained, incarcerated, recidivating within three months, and so on might be the dependent variable of interest. These proportions could be plotted for each month prior to the project, forming a pre-project time series to be compared with the post-project data. With this type of design, the evaluator cannot be certain that the pre and post populations will differ only because of the project intervention. The characteristics of the population may be changing so that persons entering earlier differ substantially from those entering later. Further, the intervention itself could alter the

characteristics of the population. In spite of these problems, the different group ITS is a useful design in many contexts.

4. Individual-Level Interrupted Time Series: In a sense, this is not an ITS design, but is more of a multiple regression approach using an historical comparison group. Nevertheless, it is worth describing at this time. Suppose the evaluator has case-by-case data on persons who entered the court system prior to the project (including the month and year they entered) and has the same type of data on persons entering the system after a new project was implemented. "Time" can be used as an independent variable in a regression equation in order to determine whether the project altered the dependent variable of interest. The logic is that any trend in the dependent variable that was occurring during the pre-project time period would be controlled, statistically, by using the month of entry as a control variable, and the impact of the project, independent of the general trend, could be ascertained.

THREATS TO VALIDITY WITH INTERRUPTED TIME SERIES DESIGNS

Several of the common threats to validity will be discussed here in terms of whether the interrupted time series designs control for them.

History

During the pre-project time periods a variety of historical events have been occurring and are producing variability in the observations. The same would be true for the post-project observations. Because the tests of significance incorporate the extent of pre and post variability

in the data when assessing the statistical significance of project effects, many of the historical events can be controlled with this design. On the other hand, any specific historical event occurring between the pre and post series of observations that differs from those occurring in the past in terms of the direction of effect or the magnitude of effect would not be controlled. As is true with comparison group and experimental designs, the multiple time series design controls for historical events that affect both series of observations in the same way.

Maturation

An interrupted time series design that uses pre and post observations on the same group of persons who are aging, gaining experience, or maturing in other ways controls for maturation effects only if the effect of maturation on the dependent variable is linear. If so, and if linear projections are made into the post time period, then the effect of maturation is contained in the trend and is properly controlled. It often is the case, however, that the effect of maturation is not a linear function of time (or age). Offense rates of juveniles, for example, tend to increase until the youths are 14 to 16 years of age and then they decline. A time series study of a group in which the first observations were taken when the youths were 12 and the intervention occurs at about age 14 through 16 will be seriously confounded by maturation. The different group time series design (in which the pre observations are on youths who are about the same age as the groups that later enter the program) would not have a serious problem with maturation effects. The multiple time series design also would control for maturation if the comparison group is equivalent to the treatment group in terms of whatever

characteristics of maturation are presumed to influence the dependent variable.

Testing

The threats to validity introduced by testing effects are controlled by the single group time series design only if the effect of repeated testing is linear and, therefore, is incorporated into the trend line projections. If the effects "wear off" after the first or second test and these observations are not removed from the pre-project data, then the design would not control properly for their effects. The different groups time series design controls for the testing effect if it is reasonable to believe that the historical groups are influenced by these in the same way as the treatment group (e.g., the persons taking the tests are about the same age, and so on). The multiple time series design also would control for them under those same conditions.

Mortality

When the mortality problem is created by the absence of data in the post time period on certain types of persons for whom data were available in the pre-project period, then the time series design does not control for this problem.

Regression to the Mean

Interrupted time series, on the same group or area, helps to control or rule out regression to the mean only if the pre-project observations include time points with scores as extreme as those that were used to select persons or areas for the treatment and only if these occurred far enough before the project to determine the presence and/or magnitude

of regression effects. A multiple time series design in which the evaluator can find one (or preferably several) other areas that had scores as extreme as those used to select cases for the project (in the historical time periods) would be better able to control for regression effects.

Selection

Random assignment designs (time series or otherwise) are the only good way to control for selection biases. With a different group time series design, however, the evaluator could compare all pre-project historical groups or persons who meet the quantitative eligibility rules for the project with all of the persons in the post period who meet those rules--even if they were not selected into the project. This procedure is extremely conservative in assessing the effectiveness of the project unless the project handled a substantial proportion of the eligible cases. If a project takes all eligible clients, however, and the evaluator can identify those in the pre-project period who would have been eligible, then the different groups time series design would control for selection bias.

ANALYZING DATA FROM INTERRUPTED TIME SERIES DESIGNS

Among the various methods and statistics that have been used to analyze interrupted time series data are the Walker-Lev tests, analysis of covariance, ordinary least squares regression analysis, the Chow test of statistical significance, the single mood test, the double mood test,

and a series of different models based on the Box-Jenkins work which are called ARIMA (auto regressive integrated moving average).[2] Of these, only the ARIMA models constitute a fundamentally different approach to time series analysis. Before describing some of these statistics and explaining (in non-technical terms) how to use them, it should be noted that all the statistics mentioned except the ARIMA models are based on multi-variate linear regression.

Five of the significance tests can be obtained from the Walker-Lev time series computer program and these will be described first.* It should be emphasized that these statistics test for different types of changes in the data. Thus, one should not expect the results from the tests to be "consistent" for a particular time series because the tests are used for different purposes and have different interpretations.

Single Mood

The single mood test fits a linear regression line to the pre-program observations and then projects an estimate of the expected value (score) for the first time period after the program has been implemented. The difference between the predicted value and the observed value for the first time point after program implementation is evaluated by a t-test. If the difference is statistically significant (at the .05 level for example), then the conclusion is that the program had an immediate impact. This test provides no information on whether the impact was maintained or whether the pre-program trend was altered by the program.

---

*Examples of how to access and use these are contained in Part II of the Handbook. Also see the discussion of the "Bellevue Citizen Involvement in Burglary Prevention Evaluation" in Section 4B for examples of their use and the actual computer output.

Double Mood

The double mood test can be used when there are a sufficient number of post-program time points to fit a linear regression line to them. A regression line is fitted to the pre-program time points and an estimate made for the expected value of a time point that lies in between the last pre-program time point and the first post-program time point. Another regression line is fitted to the post-program data and an estimate projected backwards in time to the point that lies in between the last observation for the pre-program time period and the first observation for the post-program time period. A significant t-test indicates that there was an immediate impact from the program and the impact was maintained during the post-program time period. Under some conditions the single mood may be significant and the double mood may not be significant for the same data. Consider the hypothetical data in Figure 4. In this hypothetical case the single mood test probably would be significant, indicating an immediate impact, but the double mood would not be significant.

Walker-Lev 1 (ANCOVA 1)

The first Walker-Lev test (which is identical to the first ANCOVA test) compares the regression slope for the pre-program time period with the slope for the post-program time period.[3] If the test is significant, one can conclude that the trend observed in the pre-program months was altered significantly by the program. Given certain types of changes in the time series, the Walker-Lev 1 can be significant even though neither the single nor double mood tests would be. Consider the example

FIGURE 4

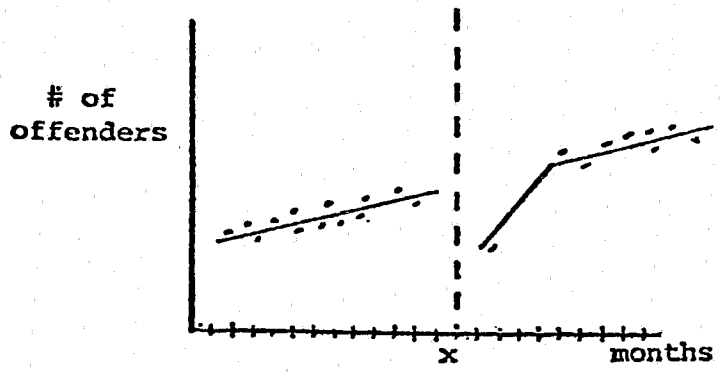


FIGURE 5

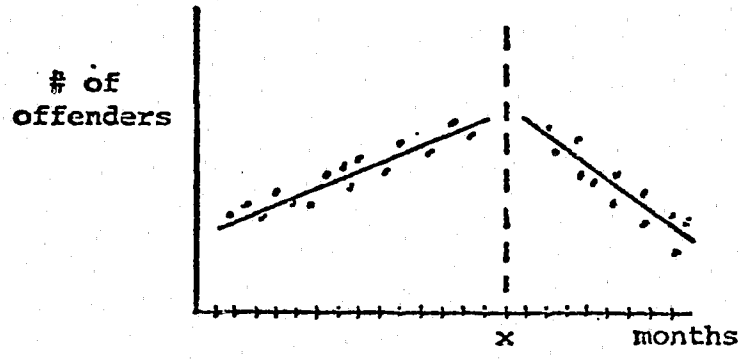
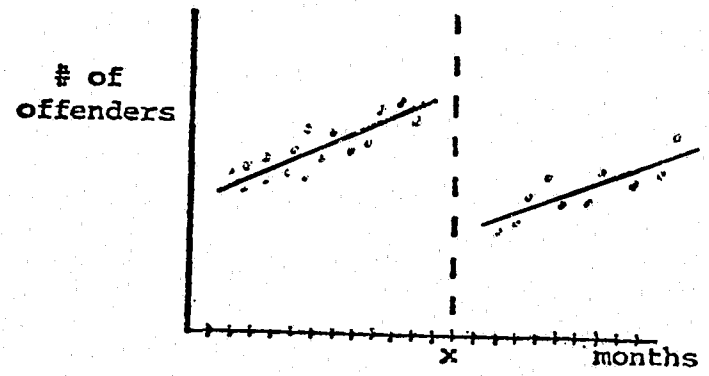


FIGURE 6



in Figure 5. In this example the program intervention clearly altered the generally upward trend of the pre-program series, resulting in a downward trend afterward. Neither the single nor the double mood tests would be significant, however, because the predicted observations would be almost identical to those observed for the month when the program began as well as for the hypothetical time point that lies in between the two series.

Walker-Lev 2 (ANCOVA 2)

The second Walker-Lev test is of interest only if the first Walker-Lev indicates that there is no difference in slopes (trend) for the pre-program time period compared with the post-program time period. If the slopes are the same (statistically insignificant differences), then the second Walker-Lev indicates whether the trend for the entire regression line is significantly different than zero.

Walker-Lev 3 (ANCOVA 3)

The third Walker-Lev test is to identify significant changes in the entire level of the series (differences in intercept values) that could be attributed to the intervention. This is done by comparing a single regression line for the pre and post data with regression lines within the pre and post that have the same slope but unique intercepts. If the first test indicated that the slopes (trend) in the pre and post data were the same, the Walker-Lev 3 is a clear test for differences in the level of the series. But if the slopes were different, according to Walker-Lev 1, then the third test is not particularly meaningful. This test is designed to show statistically significant differences for

data of the type shown in Figure 6.

Most of the statistical analysis packages for the social sciences have multiple regression programs that can be used to produce the same statistics as the Walker-Lev tests and most have analysis of covariance routines that provide the same information. The ANCOVA routine in the Statistical Package for the Social Sciences (SPSS) does not, however, provide quite all the information needed.[4] Figure 7 shows the interpretation for the results of a multiple regression time series analysis. The formula is:

$$Y = a + b_1 I + b_2 \text{TIME} + b_3 I\ \text{TIME}^*$$

where: Y = the dependent variable.

I = a dummy variable representing the intervention point; observations before the intervention would be given a score of zero; observations after the intervention would be given a a score of one.

TIME = time, measured 1, 2, 3, 4, and so on to the most recent point, using weeks, quarters, years, and so on.

I TIME = interaction between time and the intervention dummy variable (this variable is created by multiplying the score on the intervention variable [zero or one] and the score on the time variable, thereby creating a new variable for each case.

The three hypotheses tested with the Walker-Lev tests or the ANCOVA tests can be examined using various parts of the equation above, as summarized in Table 1.

The Chow test is slightly different from any of these and, in some ways, might be more useful.[5] Its purpose is to determine if the post intervention observations are from a different population than the pre intervention observations. This is done by comparing the explained

---

*For readers not familiar with regression analysis this notation is developed and explained in the following paper "Intuitive Logic of Multiple Regression Analysis."

FIGURE 7

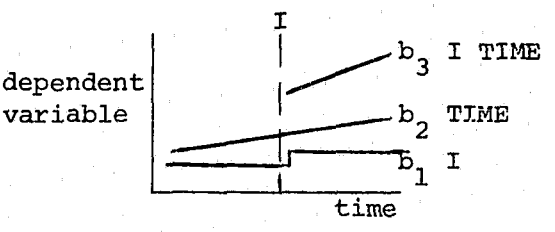
INTERPRETATION OF MULTIPLE REGRESSION STATISTICS

FOR INTERRUPTED TIME SERIES

values of the variables[1]

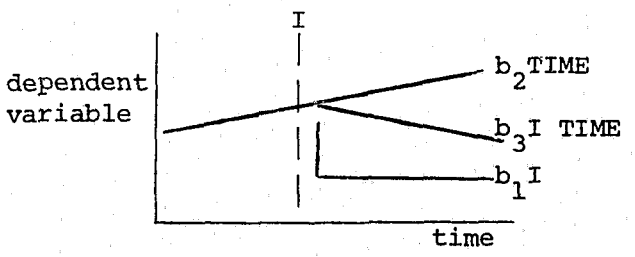
		I				
I*TIME	0	0	0	0	5	6	7	8
I	0	0	0	0	1	1	1	1
TIME	1	2	3	4	5	6	7	8

interpretation if b_1 b_2 b_3 are statistically significant and positive[2]



$$Y = a + b_1 I + b_2 TIME + b_3 I\ TIME$$

interpretation if b_1 is negative, b_2 is positive, & b_3 is negative & all are statistically significant[2]



[1] Time has integer values such as 1, 2, 3...n. I (intervention) has a score of 0 for the pre-intervention and a score of 1 for the post. The interaction term I*TIME is formed by multiplying I by time.

[2] b_3 for the interaction term indicates how the slope (trend) shifted after the intervention. b_2 for time is the pre-project trend projected into the post period, and b_1 for the intervention (I) shows the change in level of the series after the intervention.

TABLE 1. PROCEDURES FOR USING MULTIPLE REGRESSION ANALYSIS TO TEST FOR SIGNIFICANT INTERVENTION EFFECTS

HYPOTHESIS	FORMULA	EQUIVALENT TESTS	INTERPRETATION AND COMMENTS
H1. The intervention produces a change in the trend of the dependent variable.	$Y = a + b_1 I + b_2 time + b_3$ ITime	Walker Lev 1 ANCOVA 1	(a) The regression coefficient for the interaction term, b_3, gives the post-intervention slope adjustment over the pre-project data. If b_3 is significant, then the regression coefficient for time (b_2) is the pre-intervention slope (trend). (b) If b_3 is not significant, test for H2 and H3.
H2. (IF H1 is not accepted). There is an underlying trend in the dependent variable throughout the entire time period.	$Y = a + b_1 + b_2 time$	Walker Lev 2 ANCOVA 2	Note: The interaction term must be removed from the equation to test this hypothesis. The coefficient for time (b_2) gives the slope (trend) for the entire pre and post time periods. If it is significant, then the trend is different from zero.
H3. (If H1 is not accepted). The intervention produced a change in the level of the dependent variable.	$Y = a + b_1 I + b_2 time$	Walker Lev 3 ANCOVA 3	Note: The equation is the same as for H2, but as before, the interaction term must not be in the equation, even if it is not significant. The coefficient for I (b_1) gives the post-intervention intercept adjustment which can be interpreted as the magnitude of change in the level of the series.

variation obtained from a single regression line with the explained variation obtained when unique slopes and intercepts are calculated for both the pre and post time periods. Using the output from the multiple regression analysis explained above, one could calculate the value of the Chow test by conducting an F test of significance on the increase in R^2. The time variable would be entered first (and R^2 is calculated) in a step-wise regression and then the intervention variable and the interaction term entered. If the change in R^2 when all three are in the equation is statistically significant (in comparison with the R^2 when only the time variable was used) then the conclusion would be that the post intervention observations are from a different population than the pre intervention data.[6]

ASSUMPTIONS IN USING REGRESSION ANALYSIS FOR INTERRUPTED TIME SERIES

The assumptions that should be met in order to use regression/correlation analysis are described in another paper in this series and will not be examined extensively except for a major confounding problem: non-independence of the units of analysis.

A key assumption in the use of regression analysis is that the cases or units of analysis are independent of one another. In time series data the cases or units of analysis constitute time points represented by scores on the dependent variable. Since social phenomena are often <u>not</u> independent through time, this assumption may be violated. If so, the

tests of significance are inflated and the evaluator who was unaware of the problem would erroneously conclude that a significant change occurred when, in fact, it did not.

A determination of whether this assumption has been violated can be made by examining the autocorrelation of the residuals from the regression equation. If the residuals (e.g., the error in predicted values) are correlated with each other, then there is an autocorrelation problem and the tests of significance are not valid. The autocorrelation or autoregression coefficient is calculated by regressing the error at one time point with the error at the next, moving across each of the time points. In other words, the error in prediction for the first month is paired with the error for the second; the error for the second is paired with the error for the third; the error for the third is paired with the error for the fourth, and so on. The Durbin-Watson test of statistical significance for this type of error will indicate whether the autocorrelation problem is serious enough to disregard the tests of significance obtained for the various ANCOVA or Walker-Lev or regression tests.[7] (The SPSS program will give the Durbin-Watson test if that option is requested.) The Walker-Lev program does not give this statistic, but it does give the autocorrelation coefficients for Lag 1. The approximate value of the Durbin-Watson test can be calculated from the autocorrelation coefficient using the following formula:[8]

$$D \simeq 2(1-r)$$

The critical points of the Durbin-Watson test are attached to this paper as Appendix 1. If the value is greater than the upper bounds shown

(D_u) then there is no autocorrelation problem. If the value is lower than the lower bounds (D_L) then there definitely is a problem. If the value is in between the lower and upper bounds, it is not clear whether a problem exists or not.

If there is a significant autocorrelation problem, the researcher has several options available for trying to solve it. Technical discussions of these are beyond the scope of this paper, but they are described briefly and a reference given.

1. Take first differences. First differencing is done by simply subtracting the value of the dependent variable at time 1 from the value at time 2, creating a new "difference" variable. The same is done for the value at time 2 with time 3, and so on. These new values are then used with the same analysis routines described before. If the residuals are not autocorrelated, then the significance tests are accurate. Although most references on time series analysis recommend the use of first differencing to remove autocorrelation problems, some recent authors have suggested that this procedure is not a good one to use unless the autocorrelation coefficient is close to 1.0.[9]

2. Use generalized least squares. Generalized least squares is a varient of ordinary regression analysis and, when applied to time series data, involves the calculation of the autoregression coefficient and then weighting the lag one value of the variable with that coefficient. This procedure is explained by Ostrom and he also shows how it can be done using SPSS multiple regression analysis.[10]

3. Transform the data in a substantively meaningful way. The researcher might, for example, calculate rates of change between the

observation at time 1 and time 2; time 2 and time 3; and so on. This is a variant of first differencing, but one which has a substantive interpretation and might remove the autocorrelation problem. The test for intervention effects would indicate whether the rate of change in the observations was effected by the intervention.

4. Include a lagged value of the dependent variable in the regression equation. This procedure seeks to statistically control the most recent value of the observation and test for intervention effects. The ARIMA approach uses lagged values. The major problem with the technique is that the Durbin-Watson test of significance is not valid when lagged values are included and, therefore, the evaluator would not know whether the autocorrelation problem was solved with this procedure.[11]

5. Use one of the ARIMA (auto regressive integrated moving average) models. This would be the best solution to the problem, but it also is the most difficult and technically complex, since well-developed and documented statistical routines to apply these models to interrupted time series (as distinct from ordinary time series) are not generally available.

APPENDIX I

DURBIN-WATSON TEST FOR AUTOCORRELATION

Sample size = n	Pr = Probability in Lower Tail (Significance Level = x)	k = Number of Regressors (Excluding Constant)									
		1		2		3		4		5	
		D_L,	D_U	D_L,	D_U	D_L,	D_U	D_L,	D_U	D_L,	D_U
	.01	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
15	.025	.95	1.23	.83	1.40	.71	1.61	.59	1.84	.48	2.09
	.05	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
	.01	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74
20	.025	1.08	1.28	.99	1.41	.89	1.55	.79	1.70	.70	1.87
	.05	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
	.01	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
25	.025	1.18	1.34	1.10	1.43	1.02	1.54	.94	1.65	.86	1.77
	.05	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89
	.01	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
30	.025	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	.98	1.73
	.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
	.01	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
40	.025	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
	.05	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
	.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
50	.025	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
	.05	1.50	1.59	1.46	1.63	1.42	1.66	1.38	1.72	1.34	1.77
	.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
60	.025	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
	.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
	.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
80	.025	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
	.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
	.01	1.52	1.56	1.50	1.72	1.48	1.60	1.46	1.63	1.44	1.65
100	.025	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72
	.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

TABLE IX Critical Points of the Durbin-Watson Test for Autocorrelation [see equation (6-31)]

This table gives two limiting values of critical D (D_L and D_U), corresponding to the two most extreme configurations of the regressors; thus, for every possible configuration, the critical value of D will be somewhere between D_L and D_U:

FOOTNOTES

1. There is considerable debate as to whether linear predictions, used in the deterministic models, are suitable for social phenomena or whether stochastic models such as the ARIMA ones are more appropriate. See Section 2F of the Handbook, "Applications of ARIMA and ANCOVA to Interrupted Time Series Analysis."

2. See George E.P. Box and Gwilym M. Jenkins, TIME SERIES ANALYSIS: FORECASTING AND CONTROL, revised edition (Holden-Day, 1976). These models were first applied to interrupted time series by Gene Glass et al. See Gene V. Glass, Victor L. Willson, and John M. Gottman, DESIGN AND ANALYSIS OF TIME SERIES EXPERIMENTS (Colorado Associated University Press, 1975).

3. Helen M. Walker and Joseph Lev, STATISTICAL INFERENCE (Holt, Reinhart, and Winston, 1953).

4. Norman H. Nie et al., SPSS: STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES, second edition (McGraw-Hill, 1975).

5. Gregory C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regression," ECONOMETRICA, 28 (July 1969), 591-605.

6. Nie et al. present a discussion of how this can be done.

7. See Charles W. Ostrom Jr., TIME SERIES ANALYSIS: REGRESSION TECHNIQUES (Sage Publications, 1978).

8. Ronald Wannecott and Thomas Wonnacott, ECONOMETRICS (John Wiley & Sons, 1970), p. 143.

9. See Charles W. Ostrom Jr., op cit.

10. Ibid. The technique is also described in Wonnacott and Wonnacott, op cit.

11. See Ostrom, op cit.

SECTION 2D

THE INTUITIVE LOGIC OF MULTIPLE CORRELATION AND REGRESSION*

Abstract

A step by step explication of the intuitive (rather than mathematical) logic of regression analysis is provided, along with a step by step illustration of how scores for hypothetical individuals are manipulated to obtain the statistics. The procedure is then extended to the multivariate case. For evaluators who are not familiar with regression analysis, it is suggested that they calculate the hypothetical data by hand, following the instructions in the paper, in order to fully understand the logic of the calculations. The paper describes and illustrates the substantive interpretation of the correlation coefficient (r), the coefficient of determination (r^2), the regression coefficient (b), and the intercept (a).

*These materials were prepared by Anne L. Schneider and distributed to ALJE evaluators during the Model Evaluation Program.

THE INTUITIVE LOGIC OF MULTIPLE CORRELATION AND REGRESSION

There are several approaches one can take to present an "intuitive" (rather than mathematical) logic of correlation/regression analysis. The one used in this paper begins with nominal-interval data using the basic logic of analysis of variance and then proceeds to interval-interval data and the multivariate case.[1]

Introduction

Suppose that data are available on the number of arrests for 20 youths during a two-year period after they were originally arrested for burglary. It is known that 10 of the youths were in a new juvenile program (Program A) whereas the other 10 were in the traditional minimum supervision probation program (Program B). The data are arranged as shown in Table 1.

The following observations can be made:

1. There is considerable variability in the number of arrests among youths in the two programs and considerable variability in arrests even for persons within each program.
2. The average number of arrests for all 20 youths is 3.5.
3. The average number of arrests for Program A youths is 2.
4. The average number of arrests for Program B youths is 5.

Clearly, it appears that youths in Program A did better than those in Program B. It also is clear that the programs are not the only thing influencing the arrest rates of the youths. Which program a juvenile is in explains or accounts for some of the differences in number of arrests (e.g., an average of two arrests per person for Program A vs five for Program B) but being in Program A cannot explain why one youth in Program A has zero arrests, another has five, and so on. Nor can being in Program B explain why one youth

in that program has four arrests, another nine, and another eight.

One way of phrasing the relevant question is this: How much of the variability among the youths' arrest rates is accounted for (explained) by the treatment programs and how much of the variability among them is not attributable to the programs?

TABLE 1. DATA ON 20 HYPOTHETICAL YOUTHS

Program A Youths: Number of Arrests in Two Years			Program B Youths: Number of Arrests in Two Years		
1.	Sam	2	11.	George	3
2.	John	0	12.	Harvey	4
3.	Harry	3	13.	Isaac	2
4.	Wallace	1	14.	Jacob	5
5.	Albert	1	15.	Kenneth	4
6.	Alice	3	16.	Laura	5
7.	Betty	5	17.	Mary	8
8.	Carolyn	2	18.	Nancy	6
9.	Evelyn	3	19.	Paul	4
10.	Freda	0	20.	Rita	9
	Σ =	20		Σ =	50
	$\bar{Y}_A$ =	2		$\bar{Y}_B$ =	5

Grand Total = 50 + 20 = 70

Grand Mean $\bar{Y}_G$ = 70/20 = 3.5

To answer this, one must measure the total amount of variance among the group of 20 youths and then ascertain what proportion of the variance can be explained by the program they were in and what proportion is left unexplained by the program.

Partitioning the Variance

The variance is the most commonly used measure of spread (i.e., differences among the scores, variability among the scores). It is calculated by subtracting the mean score (grand mean) from each individual score to measure deviation around the mean and then squaring the results for each case. The sum of these, for all cases, is called the sum of squares.

$$\text{Sum of Squares} = (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + (Y_3 - \bar{Y})^2 + \ldots + (Y_n - \bar{Y})^2$$

Where Y_1 refers to the first case (Sam); Y_2 refers to a second case (John); and so on through the 20th case (Rita), and $\bar{Y}$ is the grand mean.

The variance is the sum of the squared deviations divided by the number of cases, but we will use the sum of squares for the subsequent examples.

The logic of partitioning the variance into that which can be explained by the program and that which cannot is as follows: We calculate the total sum of squares for the arrest data (the measure of variability for all 20 youths) and then we calculate the sum of squares that is NOT explained by which program the youth is in. As noted before, being in Program A cannot explain why the youths within the program differ among themselves in terms of their arrest scores; nor can being in Program B explain why the youths in that group differ among themselves in the number of arrests. Thus, the sum of squares (variability) within Program A is calculated and the sum of squares within Program B is calculated.* These two are added together and are called the unexplained sum of squares. Next, we calculate the amount of variability that IS explained by being in one program rather than the other.

The logic for calculating the unexplained sum of squares is easy to grasp since it is obvious that when the youths were exposed to the same program (such as Program A) but differ in arrests, the program as a whole cannot account for the differences. The logic for calculating the explained sum of

---

*See the formulas and explanation on page 4.

$TSS = USS + ESS$ (total sum of squares equals unexplained sum of squares plus explained sum of squares.)

$$TSS = (Y_1 - \bar{Y}_G)^2 + (Y_2 - \bar{Y}_G)^2 + \ldots + (Y_N - \bar{Y}_G)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}_G)^2$$

$\bar{Y}_G$ - this is the "grand" mean for all 20 youths

Y_i - rearrest score for youth 1, 2, 3, and so on until all $\underline{n}$ cases have been used.

$$USS = USS_a + USS_b$$

$$USS_a = (Y_1 - \bar{Y}_A)^2 + (Y_2 - \bar{Y}_A)^2 + \ldots + (Y_N - \bar{Y}_A)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2$$

$\bar{Y}_A$ - this is the mean number of rearrests for youths within Program A.

Y_i - rearrest score for youth 1, 2, 3, and so on until all 10 cases in Program A have been used. N = number of cases in Program A.

$$USS_b = (Y_{11} - \bar{Y}_B)^2 + (Y_{12} - \bar{Y}_B)^2 + \ldots + (Y_N - \bar{Y}_B)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}_B)^2$$

$\bar{Y}_B$ - this is the mean number of rearrest for youths within Program B.

Y_i - rearrest scores for youth 11, 12, 13, and so on until all 10 cases in Program B have been used. N = number of cases in Program B.

$$ESS = ESS_a + ESS_b$$

$$ESS_a = (\bar{Y}_A - \bar{Y}_G)^2 (N)$$

$\bar{Y}_G$ - this is the "grand" mean for all 20 youths

$\bar{Y}_A$ - this is the mean for group A N = number if Program A.

$$ESS_b = (\bar{Y}_B - \bar{Y}_G)^2 (N)$$

$\bar{Y}_G$ - this is the "grand" mean for all youths

$\bar{Y}_B$ - this is the mean for group B N = number in Program B.

The actual calculations are shown in Table 2.

TABLE 2. CALCULATION SUM OF SQUARES

Program A Youths	Number of Offenses	Total Sum of Squares $\bar{Y}_G$	$(Y_i - \bar{Y}_G)^2$	Unexplained Sum of Squares $\bar{Y}_A$	$(Y_i - \bar{Y}_A)^2$	Explained Sum of Squares $(\bar{Y}_A - \bar{Y}_G)^2$
1. Sam	2	3.5	2.25	2	0	$(2 - 3.5)^2 = 2.25$
2. John	0	3.5	12.25	2	4	$(2 - 3.5)^2 = 2.25$
3. Harry	3	3.5	.25	2	1	$(2 - 3.5)^2 = 2.25$
4. Wallace	1	3.5	6.25	2	1	$(2 - 3.5)^2 = 2.25$
5. Albert	1	3.5	6.25	2	1	$(2 - 3.5)^2 = 2.25$
6. Alice	3	3.5	.25	2	1	$(2 - 3.5)^2 = 2.25$
7. Betty	5	3.5	2.25	2	9	$(2 - 3.5)^2 = 2.25$
8. Carolyn	2	3.5	2.25	2	0	$(2 - 3.5)^2 = 2.25$
9. Evelyn	3	3.5	.25	2	1	$(2 - 3.5)^2 = 2.25$
10. Freda	0	3.5	12.25	2	4	$(2 - 3.5)^2 = 2.25$
	$\Sigma = 20$		$\Sigma = 44.50$		$\Sigma = 22$	$\Sigma = 22.50$
	$\bar{Y}_A = 2$					
Program B Youths				$\bar{Y}_B$	$(Y_i - \bar{Y})^2$	$(\bar{Y}_B - \bar{Y}_G)^2$
1. George	3	3.5	.25	5	4	$(5 - 3.5)^2 = 2.25$
2. Harvey	4	3.5	.25	5	1	$(5 - 3.5)^2 = 2.25$
3. Isaac	2	3.5	2.25	5	9	$(5 - 3.5)^2 = 2.25$
4. Jacob	5	3.5	2.25	5	0	$(5 - 3.5)^2 = 2.25$
5. Kenneth	4	3.5	.25	5	1	$(5 - 3.5)^2 = 2.25$
6. Laura	5	3.5	2.25	5	0	$(5 - 3.5)^2 = 2.25$
7. Mary	8	3.5	20.25	5	9	$(5 - 3.5)^2 = 2.25$
8. Nancy	6	3.5	6.25	5	1	$(5 - 3.5)^2 = 2.25$
9. Paula	4	3.5	.25	5	1	$(5 - 3.5)^2 = 2.25$
10. Rita	9	3.5	30.25	5	16	$(5 - 3.5)^2 = 2.25$
			$\Sigma = 64.50$		$\Sigma = 42$	$\Sigma = 22.50$

TSS = 44.50 + 64.50 = 109
USS = 22 + 42 = 64
ESS = 22.5 + 22.5 = 45

$r^2 = .41$ $r = .64$

squares is not as self evident. Program A can explain why youths in Program A differ, on the average, from all of the youths, and Program B can explain why its clients differ, on the average, from all the youths. Thus the mean of the entire 20-person group is subtracted from the mean of Group A and from the mean of Group B (once for each client) to calculate the explained sum of squares.

Having calculated the total sum of squares and the explained sum of squares, it is now possible to determine the proportion of the variance that is attributable to the differences in treatment programs. The coefficient of determination (r^2) is a summary measure of the proportion of variance explained:

$$r^2 = \frac{ESS}{TSS} = .41 \qquad r = \sqrt{.41} = .64 \text{ (correlation coefficient)}$$

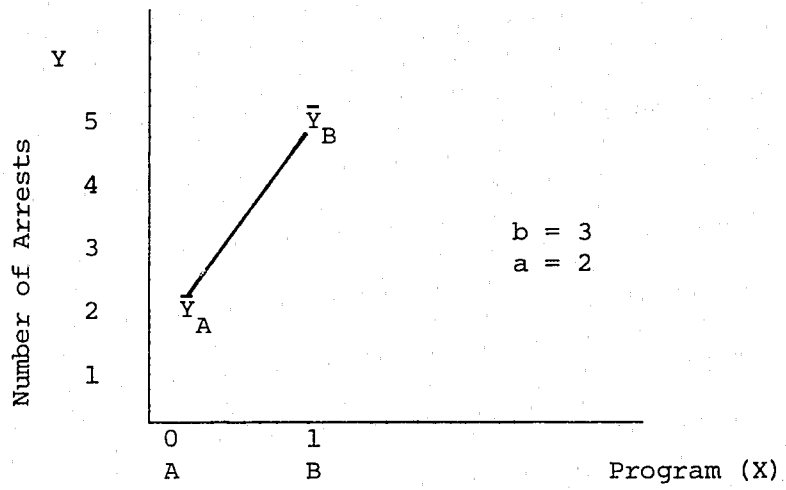
The regression coefficient (b) is simply the difference in means between the two groups. In this example (b = 5-2 = 3). If a "score" is assigned to each group (Program A = 0; Program B = 1) then the relationship can be diagrammed as shown in Figure 1. The average score for Group A is plotted above the value we have given Program A (i.e., zero) and the average score for Group B is plotted above the value we have given for Program B (1).

The regression coefficient (b) is interpreted as the change that occurs, on the average, in the dependent variable when there is a one unit change on the independent variable. Thus, in the example, when a change is made from Program A to Program B (one "unit") there is a change, on the average, from two arrests to five arrests which is a total change of three arrests.

Alpha (the intercept) is the value on the dependent variable when the independent variable is zero. Since we used zero as the score for Program A youths, alpha (in the example) will equal the mean arrest score for juveniles in Program A (two arrests).

The formula that expresses these relationships is:

FIGURE 1.



$\hat{Y}_a = 2 + 3(0) = 2$ Predicted score for each youth in Group A = 2

$\hat{Y}_b = 2 + 3(1) = 5$ Predicted score for each youth in Group B = 5

$$\hat{Y} = a + b_1 X_1$$

Where: $\hat{Y}$ is the predicted (estimated) score for a youth

a is the intercept value (2 in this example)

b is the regression coefficient (which has a value of 3 in this example

X is the independent variable (which in this example, is scored as zero if the youth is in Program A and as one if in Program B).

The correlation coefficient (.64) is a summary measure of the spread or degree of accuracy that is observed when one actually "predicts" the number of arrests for each youth. The coefficient of determination (r^2) is the proportion of variance in arrests explained by the program variable.

Extension to Multiple Regression

As noted previously, there must be some other variables influencing arrest of these youths, since only part of the differences among them is

attributable to being in one program rather than the other.

Examination of the original data in Table 1 indicates that sex may account for some of the variability in scores. Of particular concern, however, is whether the variance presumably explained by the program variable actually is due to the program effects or whether it is partly due to differences attributable to sex.

In order to find out, one could redo the analysis by first letting the sex variable explain all of the variance in arrests that it can and then calculate how much of the residual (left-over) sum of squares can be explained by the program variable.

This procedure is illustrated in Table 3 using cases from the original data. The best prediction of arrests for boys in either program would be the average number of arrests for all boys, and the best prediction for girls would be the mean number of arrests for all the girls. This becomes the "predicted" or "expected" score on arrests for each case and it is subtracted from the actual (original) score producing the residual variance that cannot be explained by sex.

For example, the mean number of arrests for boys in this analysis ($\bar{Y}_m$) is 2.5. Thus one would expect Sam to have 2.5 arrests; John should have 2.5 arrests; and so on. Sam, however, only has two arrests and therefore has .5 fewer arrests than were expected. John had no subsequent arrests and therefore had 2.5 fewer arrests than expected. The residual scores represent the variance in arrests that cannot be explained by the youth's sex and these scores become a new dependent variable. One then proceeds to calculate the proportion of the variance in these residual scores that can be explained by the program variable. Thus, a new total sum of squares is calculated for the residual scores, a new unexplained sum of squares is calculated, and a new

TABLE 3. CALCULATING THE RESIDUALS

	MALE				FEMALE		
Program A	Y	$\bar{Y}_G$	Residual Score		Y	$\bar{Y}_G$	Residual Score
1. Sam	2	2.5	- .5	6. Alice	3	4.5	-1.5
2. John	0	2.5	-2.5	7. Betty	5	4.5	.5
3. Henry	3	2.5	.5	8. Carolyn	2	4.5	-2.5
4. Wallace	1	2.5	-1.5	9. Evelyn	3	4.5	-1.5
5. Albert	1	2.5	-1.5	10. Freda	0	4.5	-4.5
$\Sigma = 7$	$\bar{Y}_{MA} = 1.4$			$\Sigma = 13$	$\bar{Y}_{FA} = 2.6$		
Program B							
11. George	3	2.5	.5	16. Laura	5	4.5	.5
12. Harvey	4	2.5	1.5	17. Mary	8	4.5	3.5
13. Isaac	2	2.5	- .5	18. Nancy	6	4.5	1.5
14. Jacob	5	2.5	2.5	19. Paula	4	4.5	- .5
15. Kenneth	4	2.5	1.5	20. Rita	9	4.5	-4.5
$\Sigma = 18$	$\bar{Y}_{MB} = 3.6$			$\Sigma = 32$	$\bar{Y}_{FB} = 6.40$		
$\bar{Y}_M = 2.5$				$\bar{Y}_F = 4.5$			

explained sum of squares is calculated.

The r^2 that is produced from the new analysis is a partial coefficient of determination (the new r value is a partial correlation coefficient). The new regression coefficient is a partial regression coefficient. Substantively, r^2 represents the proportion of the variance in arrests that can be explained by which program the youth was in after having statistically controlled for the effect of sex. The partial regression coefficient represents the amount of change in arrest scores for each unit of change in the program variable after sex has been statistically controlled.

The more familiar formula for multiple regression is:

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

(Note: most existing calculation routines do not actually let one variable explain all the variance it can, calculate the residual, and then let another variable explain as much of the residual as it can. Instead, the calculation routines partial the explained variance to X_1, $X_2 \ldots X_N$ simultaneously. Nevertheless, the intuitive logic of multiple correlation is easy to understand using the step by step procedure described above.)

Extension to Interval Level Data

The extension of the logic to interval data for both the independent and the dependent variables is quite straightforward. Figure 2 shows the relationship between parental income (X) and the number of self-reported offenses(Y) for a hypothetical group of 12 ninth grade boys. Three of the boys have parents with incomes of $5,000. These youths reported a total of 33 offenses, for a mean of 11. The three boys whose parents have incomes of $10,000 reported a total of 30 offenses, for a mean of 10. The score (on the dependent variable) for the subset of persons who have a common score on the independent variable is plotted on the diagram.

The regression line is calculated in such a way as to maximize the proportion of the total sum of squares that can be explained by the independent variable which also means that the regression line minimizes the squared deviations around the line (the unexplained sum of squares is minimized). This type of regression analysis often is referred to as least squares regression.

Figure 3 has been drawn to illustrate the interpretation of the statistics that one obtains from correlation/regression analysis.

FIGURE 2.

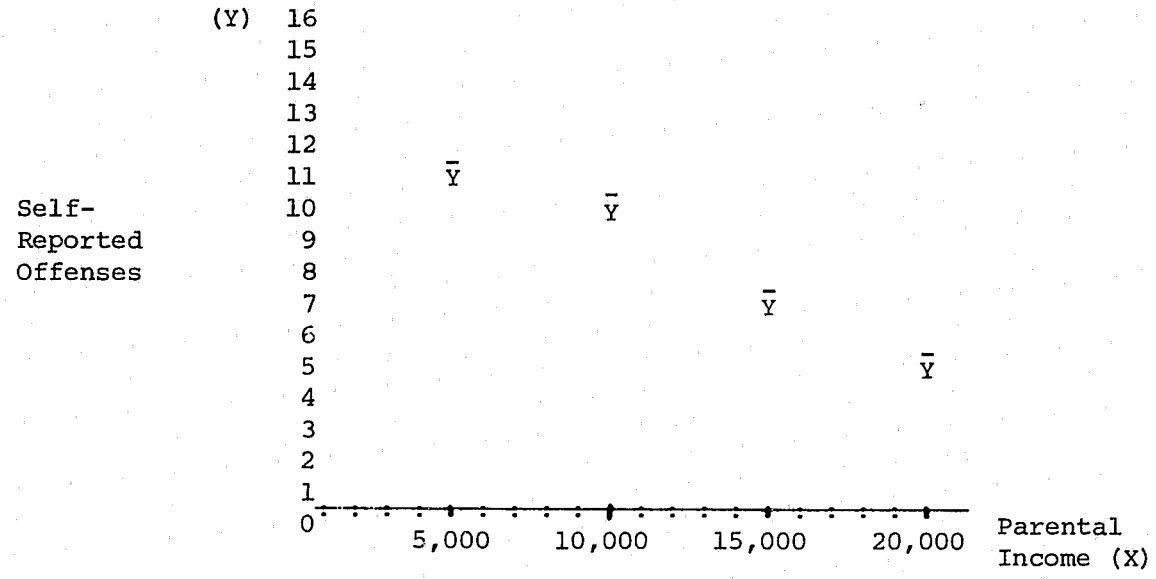
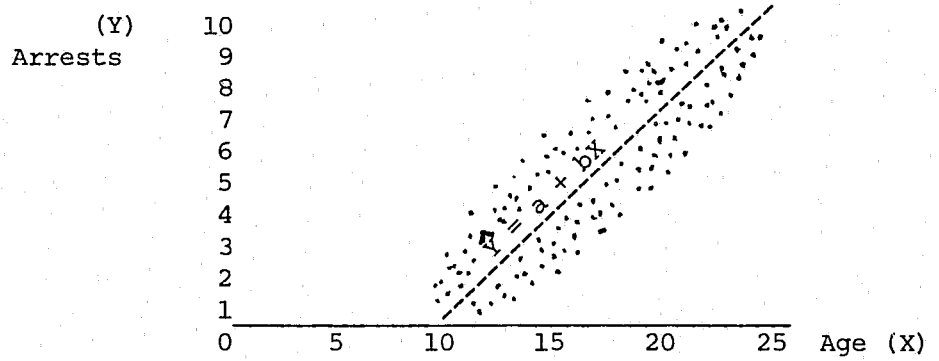


FIGURE 3. REGRESSION DIAGRAM, INTERVAL DATA



b - (regression coefficient, slope). Each unit of change in the independent variable is associated with b units of change in the dependent variable. In the example, b is .5 indicating that each unit of change in age (i.e., each additional year) is associated with the addition of one-half an offense. It takes a change of two years in age to produce, on the average, one additional offense.

beta (standardized regression coefficient). A change of one standard deviation on the independent variable is associated with a change of one standard deviation on the dependent variable when beta equals one.

a - alpha (intercept). The value of Y (dependent variable) when X is equal to zero. In the example, the regression line crosses X at 10 years,

indicating that on the average there were no offenses committed by youths with an age of 10 and that if age were theoretically zero, there would be a -8 offenses.

r^2- (coefficient of determination). The proportion of variance in the dependent variable explained by the independent variable. In the example, the actual scores are shown as points on the graph. The sum of the distance between each point and the regression line, squared, is the unexplained sum of squares.

$1-r^2$ - proportion of variance in dependent variable not explained by the independent variable.

r - correlation coefficient. The correlation coefficient has no straightforward interpretation and is most easily understood by squaring the value to obtain the proportion of variance explained by the independent variable.

Application to Evaluation

There are many applications of correlation and regression analysis to evaluation research. Two will be illustrated here.

Situation 1. Juveniles have been randomly assigned to Program A and to Program B.

As shown in the first example (Table 1), correlation analysis can be used to calculate the proportion of variance on the dependent variable attributable to the program. This can be done with an analysis of variance routine or with any standard correlation/regression program. The formula, for this type of regression is:

$\hat{Y} = a + b\ X$

Y - the dependent variable

a - the intercept (alpha)

b - the regression coefficient (slope)

X - the independent variable which, in this example, has a score of zero if the youth is in Program A and a score of one if the youth is in Program B.

Output from the program will include the value of the correlation coefficient (r), both the standardized (β), and unstandardized (b) regression coef-

ficients, the intercept value (called alpha, a, or the constant), and for most computer routines the output will show the "within group" sum of squares (i.e., unexplained variance) and the "between group" sum of squares (explained variance).

Situation 2. A jurisdiction introduced a new program (Program A) in 1977 for juvenile felony offenders and data have been collected about youths in the program. Data also have been obtained on a group of juvenile felony offenders from 1976 to be used as a comparison group (Program B). Program A clients constituted only a fraction of the 1977 youths and there is no information on the criteria used to determine which juveniles would be in Program A. It is suspected that youths in the new program may be younger and have fewer prior offenses than the total population. Therefore, the investigator wishes to ascertain whether Program A clients have lower recidivism scores than Program B which are due to the program independently from differences in recidivism that might be produced by differences between the age of youths in the two groups or differences in the number of prior offenses.

The formula is as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

Y - dependent variable (recidivism)

a - alpha (intercept)

b - partial regression coefficients

X_1 treatment variable (Program A=1; Program B=0)

X_2 age

X_3 number of prior offenses

In addition to the output described in Situation 1, the output from the program will include R^2 (multiple coefficients of determination) which represents the amount of variance in Y that can be explained by all of the variables in the equation. Of major concern is whether the treatment variable has a

statistically significant effect on Y when the other variables (such as age and prior contacts) have been statistically controlled. Most computer programs will provide a significance test for each of the partial regression coefficients. If the one for the treatment variable is significant, then the program apparently had an impact on the dependent variable that is independent of differences between the group in terms of age, race, sex, prior contacts, and the other characteristics of the youths that were used as control variables in the equation.

A note of caution. If the researcher suspects that the "easy" cases were generally assigned to Program A (or the "hard" cases were generally assigned to Program A) then statistical controls, using multiple regression, will not entirely remove variance in Y attributable to pre-treatment differences in the youths. If the "easy" cases are in Program A, there is a danger that the treatment variable will appear to be effective when, in fact, it is not. If the "hard" cases tend to be in Program A, there is a danger that the treatment will appear to have no impact or to even be "worse" than the pre-program group when, in fact, it is not. Although multiple regression analysis is one of the better analysis techniques (perhaps the best) for non-equivalent comparison group designs, its use does not relieve the evaluator of the responsibility to identify alternative explanations for the results obtained with the regression analysis.

Problems & Issues in the Use of Regression & Correlation Analysis

There are several assumptions which should be met before one uses regression or correlation analysis.[2] Some of these are more important than others in that violation of certain assumptions will not have much impact on the conclusions that are drawn. In this section we will discuss some of the assumptions which, if violated, can have serious consequences for the researcher's ability to draw accurate conclusions from the data.

1. Assumption of Linearity

Correlation-regression analysis assumes that the relationship between the dependent variable and the independent variable is linear, as shown in Figure 3a below. If the relationship is not linear, such as shown in Figure 3b, the regression equation and correlation coefficient may not provide accurate descriptions of the relationship.[3]

Figure 3a: Linear

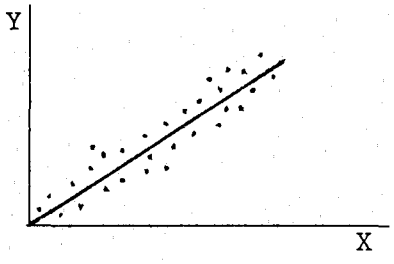
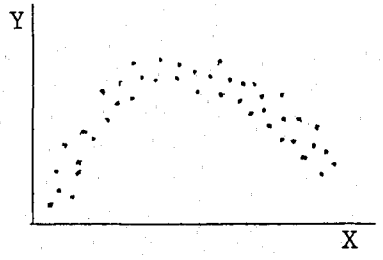


Figure 3b: Non-Linear



The investigator's theory should specify whether the expected relationship is linear or, if not, what type of relationship is expected. The researcher always should examine a scatterplot of the relationship between the dependent variable and each independent variable to be used in the analysis before proceeding to multiple regression analysis. If the relationship is not expected to be linear, from a theoretical perspective, and is not linear,

based upon examination of the scatterplot, then the data and/or the equations should be transformed and a non-linear analysis conducted.*

2. Assumption of Normal Distribution

Variables used in regression-correlation analysis should be normally distributed. In practice, however, data (especially from small samples) are rarely perfectly normally distributed. There are two problems which the researcher should always examine before proceeding with regression or correlation analysis. The first is to determine whether there are any extreme outliers in the data. An outlier is a case which has a value much higher or much lower than any of the other cases in the study. For example, one might measure the number of prior offenses and determine that 99 percent of the cases have six or fewer prior offenses but one individual has 25 priors. This is an extreme outlier.

If the data are otherwise normally distributed, the simplest solution is to reduce the value of the outlier to a number just larger than the next highest number. Like any transformation, this preserves the rank ordering of the cases. In other words, one would group seven or more priors into a single category in order to "normalize" the distribution.

The second problem that often is encountered involves a highly skewed variable in which most of the cases have either a high or low score and the others take on a considerable range. For example, one might find that 90 percent of the cases have a score of zero on number of prior offenses and the remainder have scores ranging from one to 10 priors. There are no simple solutions to this type of skewness problem, but the investigator should be aware of the fact that the correlation coefficient cannot reach its maximum value (of +1 or -1) when the variables are badly skewed in opposite directions.

---

*See "Prediction Models," Section 2E of this Handbook for a discussion of this.

Thus, the statistics one obtains when using a skewed variable are more conservative than they would be if the data were normally distributed.

3. Independence of Observations

The assumption of independent observations means that each unit of analysis has a score which is independent of the other units. For example, it is assumed that one person's score on recidivism is not influenced by another person's score. This assumption is most likely to be violated when one has measured both the independent and dependent variable at several different time points. For example, one might correlate the number of arrests with the number of convictions using monthly data. Thus, the first unit of analysis might be January, which has 20 arrests and five convictions; the next is February, which has 25 arrests and 10 convictions; and so on. These units of analysis (months) probably are not independent of one another due to trends in both of the variables.

Violation of the assumption of independence does not affect the regression coefficient nor the correlation coefficient, but it does result in inflated F and t values used in testing the significance of the coefficients.

4. Interval-Level Measurement

One of the assumptions of regression-correlation analysis is that all of the variables have been measured at the interval level. Often, there is confusion over what this means and there is considerable disagreement concerning how important it is. It is generally recognized that one can use categorical variables in regression analysis if one has scored the variables properly. (This sometimes is called dummy variable regression analysis.) For example, suppose one is comparing the effectiveness of three different treatment programs in relation to recidivism rates. In order to conduct a multivariate analysis, three variables should be developed. Persons who are

in Program A are given a score of one on the variable representing Program A; others are given a score of zero. Persons in Program B are given a score of one for the Program B variable; others are given a zero on Program B. Persons who are in Program C are given a score of one on the Program C variable, and others have a zero on this variable. Scores of zero and one are often used, but other numbers are permissible. When conducting the analysis, however, one of the dummy variables must be omitted from the equation because scores on it are completely determined by scores on the other two dummy (categorical) variables. The omitted category is called the reference category and predicted scores for it are given by the intercept value (alpha). In normal regression analysis, the value of alpha is interpreted as the value of Y (the dependent variable) when X is zero. In dummy variable regression, the value of alpha shows the expected (predicted) score on Y when an individual has a score of one on the reference category.

It is critically important to remember that when categorical (nominal) variables are used in regression analysis, one must not give scores of one, two, three, four, and so on, to the different categories, since this type of scoring presumes some kind of underlying metric order across the categories. Nominal and categorical variables, by definition, do not have any underying metric order except "presence" or "absence" in the category.

Another issue in regression analysis concerns the use of ordinal data where one has a variable with three or more scores (one, two, three, for example), but the "true" distance between a score of one and a score of two is not equal to the "true" difference between a score of two and a score of three. There is considerable disagreement concerning the consequences of using regression-correlation analysis on ordinal-level variables, but the best information, at this time, is that the major consequence is usually

one of depressing the magnitude of the regression coefficient and correlation coefficient.[4]

5. Cases to Variables Ratio

One of the most commonly overlooked problems in multiple regression analysis is that the number of independent variables used in the analysis should not exceed one for approximately every 15 cases. For example, if the investigator has 50 cases in the analysis, no more than three independent variables should be used in the multivariate analysis. If this ratio is exceeded, the F value of the multiple coefficient of determination (R^2) will begin to drop and the substantive interpretation of the results can become quite meaningless.[5]

FOOTNOTES

1. Regression analysis is discussed in most standard statistical texts. The bibliography in Section 6 of the Handbook lists several texts that include such discussions.

2. An excellent discussion of the assumptions and their relevance is found in Eric A. Hanushek and John E. Jackson, STATISTICAL METHODS FOR SOCIAL SCIENTIST (Academic Press, 1977).

3. There are many other kinds of non-linearity.

4. See Brent Rutherford, "The Accuracy, Robustness, and Relationships Among Correlational Models for Social Analysis," presented at the annual meeting of the American Political Science Association in 1972.

5. Robyn M. Dawes and Bernard Corrigan, "Linear Models in Decision Making," PSYCHOLOGICAL BULLETIN, 81 (1974), 95-106.

SECTION 2E

PREDICTION METHODS*

Abstract

This paper deals with the problems encountered when an experiment or quasi-experiment is not possible. In these cases the evaluator must create some sort of predicted outcome against which the program treatment can be compared. Actuarial tables and prediction models are discussed as possible options. It is argued that actuarial tables have a number of problems that may make them inappropriate in most criminal justice evaluations. Alternatively, two prediction model techniques are considered: Multiple classification analysis and multiple regression analysis. Multiple classification analysis offers an approach that attempts to reflect the full detail of the sample data. However, under certain conditions it is possible to disregard much of this detail and make accurate predictions employing the multiple regression model.

---

*This paper was written by Jerry Medler, based on his presentation and that of Robyn Dawes at an ALJE special forum.

PREDICTION METHODS

The Problem

In many evaluation contexts it is not feasible to create a meaningful control group or even a comparison group which has not been exposed to the program treatment. This generally comes about because of the inability to make random assignments to experimental and control groups from a designated pool of treatment clients. Because the treatment pool is often a subset of a larger population (i.e., selected as a target population), descriptive statistics for the larger population may be available but inappropriate for benchmarks or comparisons for evaluating the effects of the program. This lack of an appropriate comparison group leads to the need for a prediction of how the treated subgroup would have behaved without treatment. Such an estimate could then be compared to the actual observed behavior after treatment. A test of the significance of any observed difference could then serve as the inferential basis for evaluation of treatment effects.

It must be recognized, however, that a prediction for "no treatment" behavior is vulnerable to the usual validity problems of quasi-experiments. It is still possible that historical events or other factors external to the program could cause a shift in behavior after the program is underway which would be mistaken for program effects. In short, a prediction method is not a substitute for a control group.

The Actuarial Approach

There are many different approaches to making predictions. For our purposes we will distinguish between actuarial tables and what we will call prediction models. An actuarial table is best thought of as an n-variable contingency table. The variables themselves are frequently demographic characteristics such as age, race, and sex. Cross tabulation of the set of variables creates a large (often very large) number of categories. The set of categories is then cross tabulated with the behavior we are trying to predict, such as recidivism. The relative frequency of the behavior is used as a direct estimate of the probability of the occurrence of the behavior. For example, if we examine the category of white-male-sixteen and find that 25 percent of them recidivate in six months, we infer the probability of six month recidivism is .25. To predict the total recidivism for a client pool (say, males less than sixteen years old), we would select the relevant subset of cells from the actuarial table (e.g., all cells in which there are males sixteen or younger). We would then break down the client pool by the same variables used to construct the actuarial table. Because the client pool is a subset of the population for which the actuarial table was compiled there will be fewer cells in the client pool table than in the actuarial table. However, for each cell in the client pool table, there should be a corresponding cell or probability estimate in the actuarial table.

The information in the two tables can be combined to make a prediction. Let N_{ijk} represent the number of subjects in a given cell of the client table, where i stands for the i^{th} category of variable A,

j for the j^{th} category of variable B, and k for the k^{th} category of variable C. Then P_{ijk} represents the probability in each category of the actuarial table. $P_{ijk} \cdot N_{jik} = E_{ijk}$, producing a predicted number of recidivists for each cell of the client pool table. The sum, $\Sigma E_{ijk} = R_{total}$, gives the total number of recidivists expected in the treatment group. This sum then becomes the benchmark against which we evaluate the effectiveness of the program under study.

At first glance this approach seems very useful: We have a prediction of untreated behavior without the expense of a control group. In addition, we can analyze the predictions and actual behavior on a cell-by-cell basis which could give us additional information about the relative success of the program for discrete categories of clients. For example, we might find that there are fewer recidivists than predicted for those clients fourteen and under, while those over fourteen are recidivating at approximately the predicted rate. Such a finding might be very useful on either redefining the client pool or adjusting the treatment. When actuarial tables are viewed as an analysis of variance they seem particularly appealing. As more cells are added to an actuarial table (by adding variables or adding categories to the variables) the variation within cells will generally go down. Thus, the larger the table (the greater the number of cells) the lower the unexplained sum of squares. Given a set of variables and a set of categories for the variables, actuarial tables reduce the unexplained variation to a minimum. Viewing the actuarial table as the focus of analysis (as distinct from a prediction tool) we have reached the upper bound of our explanatory power with the set of variables used to create

the table.

In practice there are several difficulties with using actuarial tables. The most obvious of these problems is an empty cell in the actuarial table. When the tables are constructed from several variables with several categories, it is easy to generate thousands of cells. Even with thousands of cases in the actuarial table we may not have any female/black/eleven-year-old/bad check writers. If such a person ends up in the client pool we are hard pressed to make a prediction about her recidivism, as we have no basis for a prediction. A slightly less severe version of the empty cell problem is the near-empty cell. Here we need to recognize that the observed relative frequency is only an estimate of the probability of behavior. More importantly, this estimate can be wrong. The standard error of a proportion is given by $\sqrt{\frac{p \cdot q}{N}}$. Exploration of this error leads to the realization that the number of cases (N) is very important. For example, where p=.25 the standard error of the estimated proportion is 4% for 100 cases, but nearly 14% for ten cases. This demonstrates the problem with predictions based on relatively small numbers of cases--they can have potentially large errors. This reveals the dilemma of actuarial tables. If we increase the number of cells, we reduce the unexplained variation to a minimum, which seems desirable. However, we are simultaneously maximizing the standard error of the predictions in the table, which is not desirable.

Theoretically the problems of the actuarial table could be handled by merely adding enough cases to assure adequate cell size or, alternatively, acceptable standard errors. However, in practice this does

not seem to be an acceptable approach. For example, in a small jurisdiction there may be too few cases to fill an actuarial table. It would be possible to expand the number of cases in the actuarial table by going back in time to include more cases. This, however, raises a serious question of validity.

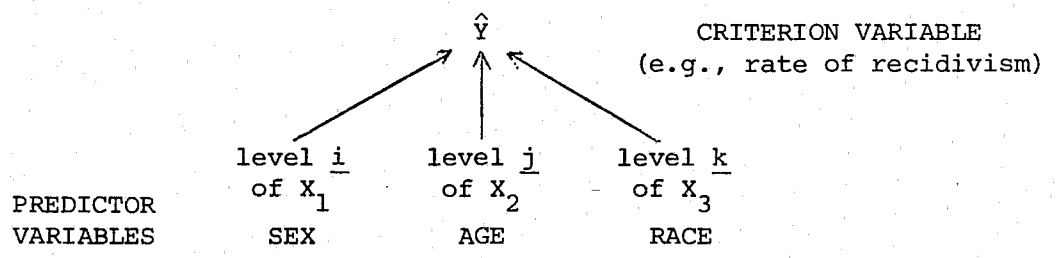
The popularity of recreational drug use and the response of the criminal justice system over time can illustrate the problem. Many jurisdictions vigorously prosecuted drug users but later (for a variety of reasons) lessened their activity in this area. Consequently, an actuarial table based on vigorous prosecution is of little value in an era of more relaxed control. More to the point, changes in the direction (less vigorous prosecution) will overstate the effect of a program. Alternatively, changes in the opposite direction (more vigorous prosecution) will lead to an understatement of the program effects. At the very least, actuarial tables based on extended periods of time would have to be examined for trends and, if present, adjustments made for extrapolation to the program period. This, however, vastly complicates the use of actuarial tables and perhaps undermines much of their attraction as a simple-to-use tool.

The Model Building Approach

Because some of the problems with the actuarial approach may be difficult (if not impossible) to overcome in particular evaluation contexts, it is important to explore some alternatives. The major alternative is to construct a model of the behavior we want to predict. The prediction model, like the actuarial table, can then be used to generate

a prediction of untreated behavior, a benchmark for evaluating the effects of a particular program. In general, prediction models are better able to make use of relatively small numbers of cases and thereby avoid some of the problems of empty or near-empty cells in an actuarial table. In the simplest terms, this is accomplished by averaging the information in the cells of the actuarial table to summarize the effects of a predictor variable. In doing this we often overlook, or smooth out, irregularities in the actuarial table in hopes of gaining a better picture of the overall predictive role of a variable.

The basic tasks of the prediction model can be diagrammed as follows:



The goal is to predict an event such as recidivism for an individual or group of individuals who are characterized by their position on a set of predictor variables. In the case of recidivism, the prediction $\hat{Y}_{ijk}$ takes the form of a proportion of a group or a probability for a single individual.

Without experimental control we are almost assured that predictor variables will be correlated. Prediction models therefore must be able to "sort out" the partial effects of interrelated predictors. The raw or unadjusted strength of a relation can be estimated by a bivariate regression or by calculating the explained sums of squares associated

with the categories comprising a variable. In contrast, adjusted strengths are expressed by weights, such as partial regression coefficients or adjusted sums of squares. In the two models to be discussed this adjustment is accomplished by solving a set of simultaneous equations (called normal equations) which take into account the correlations among predictors and express the importance of each predictor "holding constant" all the other predictors.

There are several decisions that need to be made in constructing a prediction model. Perhaps the most basic question is the shape of the relation between the predictor variables and the criterion variable. The simplest form is a linear relation. However, actual data often suggest curves. Once the linear form is abandoned there are many options. Figure 1 contrasts the simpler curves of power functions and logarithmic functions with the straight line. All of the curves in Figure 1 can be considered conditional monotones; as $\underline{x}$ increases $\underline{y}$ increases. It is conceivable that more complex shapes could be suggested by the data. Figure 2 illustrates possible non-monotones. If the data display non-linear patterns, a better fitting model generally can be specified by selecting an appropriate exponent for the values of the predictor variables. For instance, a curve might be modeled by $\hat{Y} = X^2$ or $\hat{Y} = \log X$.

In addition, multivariate prediction models must specify how the effects of the predictor variables combine. In general, there are two options available. The simplest is to combine effects by addition--an additive model: $\hat{Y} = X_1 + X_2 + X_3$. However, under some conditions it can be argued that multiplication is more appropriate--a multiplicative

FIGURE 1

VARIOUS ILLUSTRATIVE FORMS OF CONDITIONALLY MONOTONE RELATIONS

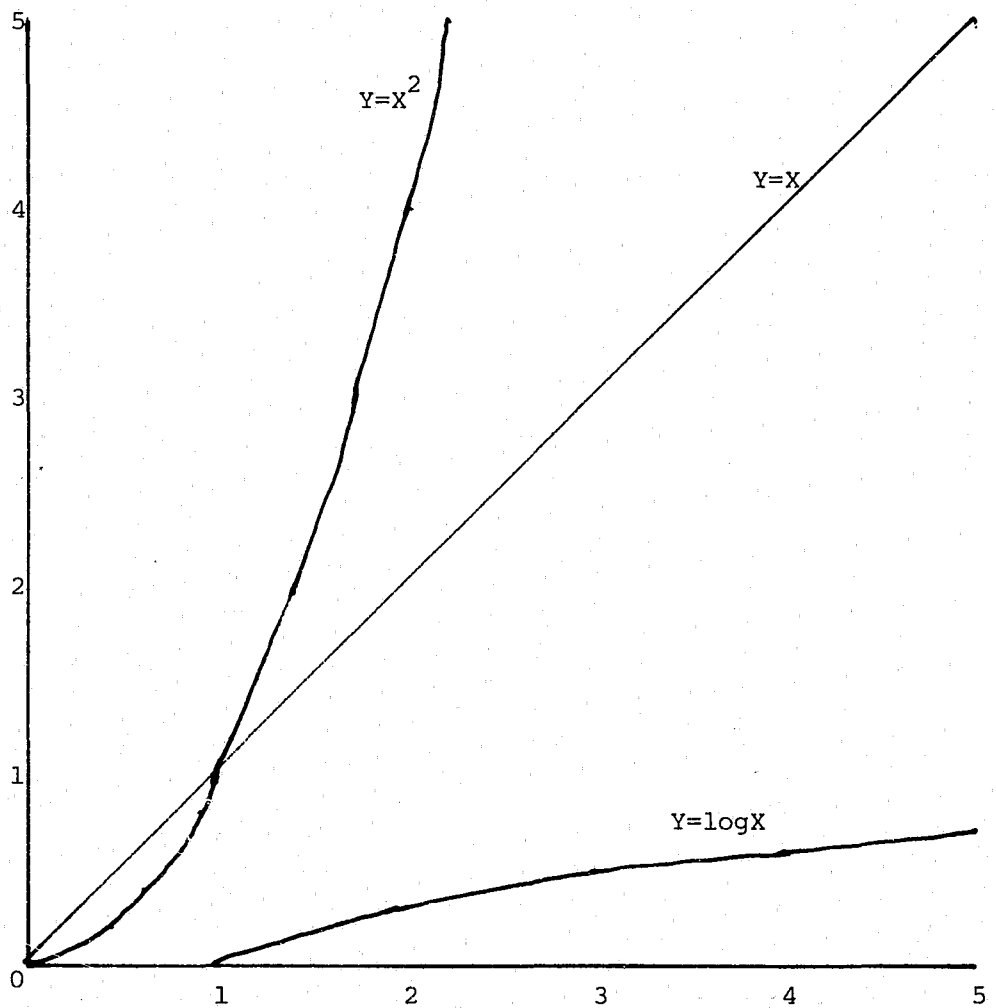
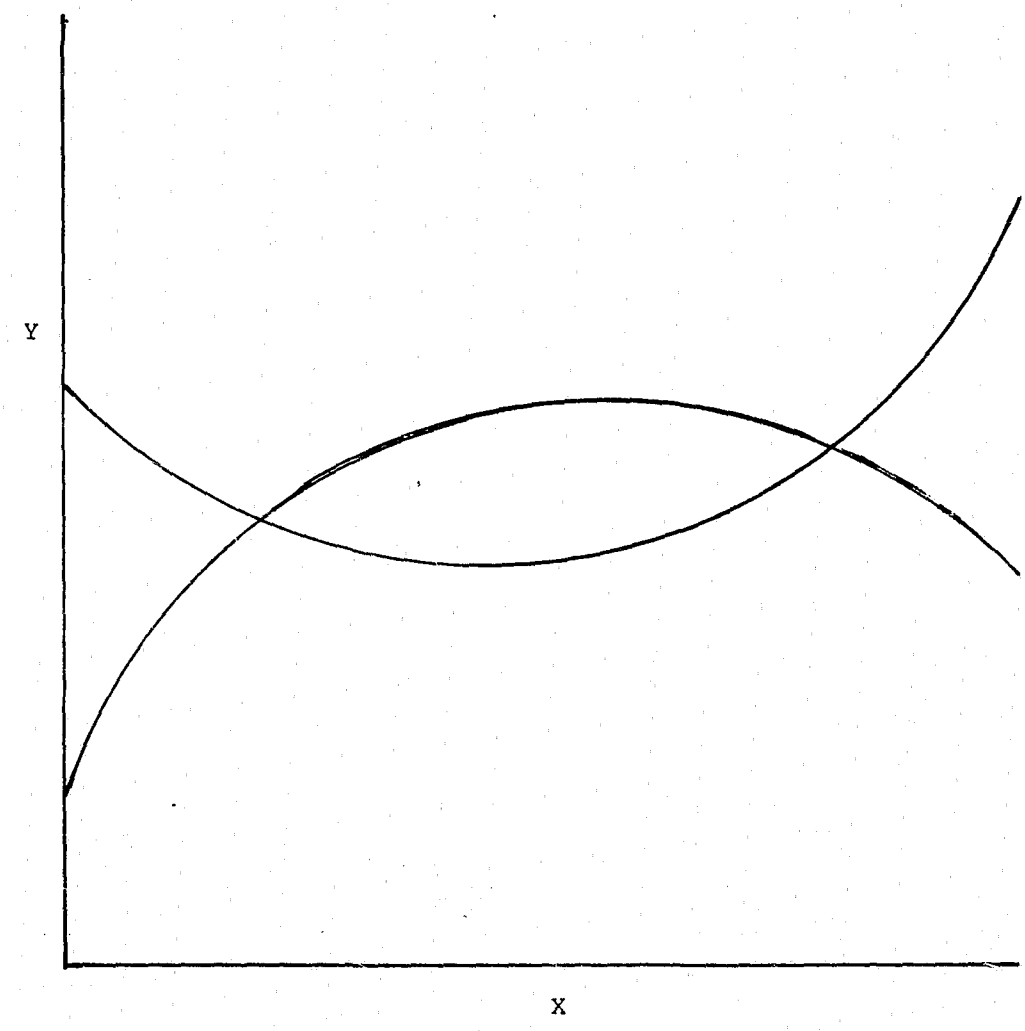


FIGURE 2

ILLUSTRATIONS OF CURVES THAT ARE NOT CONDITIONALLY MONOTONIC



model: $\hat{Y} = X_1 \cdot X_2 \cdot X_3$. It is possible that addition and multiplication may be mixed: $\hat{Y} = X_1 + X_2 \cdot X_3$. Usually social theory does not specify the shape of the relation between predictors nor whether the method of combination is additive or multiplicative. We are left to decide these matters inductively by examining the data.

In practice the data are seldom clear. The basic strategy is to fit alternative models and measure how well one model fits compared to another. This raises the question of how to measure "goodness of fit." The generally accepted principle is to minimize the error (e) between the predictions of the model ($\hat{Y}$) and the observed values of the criterion variable (Y). The difference, $e=(Y-\hat{Y})$, can of course be positive or negative, so it is squared, $e^2=(Y-\hat{Y})^2$. This squared error is then averaged for all predictions (all subjects in the sample) as: $\frac{\Sigma e^2}{n} = \frac{\Sigma(Y-\hat{Y})^2}{N}$. It is this mean summed square error (MSE) that prediction models seek to minimize. The best model is that model which produces the smallest mean squared error. However, there can be tradeoffs involved. By complicating the model with multiplicative terms and a variety of exponents, we may be able to marginally reduce the MSE. However, the price we pay may be a vastly complicated model.

Multiple Classification Analysis

Perhaps the most flexible approach to prediction models is based on an extension of analysis of variance called multiple classification analysis (MCA).[1] This approach is closely related to the use of actuarial tables and serves as a good example of how a prediction model may be derived from categorical configurations such as contingency tables.

The goal of the model can be stated as predicting the mean value (or proportion) of some criterion variable when the subject is in a particular category or cell of a contingency table. The flexibility of MCA lies in the relative lack of assumptions which need to be made about the data. For example, MCA is able to predict proportions such as the percentage of clients who might recidivate from unordered categories such as sex (male, female), as well as from ordered categories such as age (12, 13, 14, 15...). Moreover, MCA does not assume the effect of an ordered variable is linear. This is particularly attractive because it allows us to model U-shaped or inverted U-shaped relations. In predicting recidivism this is valuable because the rate may rise and then fall with age.

The major limiting assumption of MCA is that the effects of the predictor variables are additive. However, the assumption of additive effects may not be appropriate. In this case, interaction terms or multiplicative effects must be explicitly introduced if the model is to reach its maximum predictive power. This problem is discussed below.

The general model for MCA can be written:

$$Y_{ijk} = \bar{Y} + a_i + b_j + c_k + \ldots + e_{ijk}$$

where Y_{ijk} is the predicted value of the criterion variable for the subjects in the $\underline{i}^{th}$ category of $\underline{A}$, the $\underline{j}^{th}$ category of $\underline{B}$, and the $\underline{k}^{th}$ category of $\underline{C}$. The prediction is defined by references to the mean of the criterion variable $\bar{Y}$. Coefficients (a_i, b_j, c_k...) are added or subtracted according to the position of the subject in the cross

classification. The term e_{ijk} is an error term which represents the deviation of each individual from the predicted value for that category of the cross classification.

Multiple classification analysis creates adjusted coefficients for each variable which state the net effect of a particular category "holding constant" the effects of all other predictor variables. With these adjusted coefficients it is possible to visualize the effects of the category of each variable. This can be accomplished by plotting the values (a_1, a_2, a_3...). If the categories of the variable are ordered, as with age, the plot of the coefficients reveals the presence or lack of linearity in relation with the criterion variable. It should be noted that in solving the set of normal equations MCA sets the grand mean, $\bar{Y}$, equal to zero. Thus, the coefficients of each category (a_j, b_k...) are stated as positive or negative deviations from the grand mean.

The overall explanatory power of each variable is also calculated by MCA. Coefficients with and without adjustment for intercorrelated predictors are available and serve to summarize the strength of each variable in determining the criterion variable. The overall predictive power of the model is also expressed by a summary coefficient that can be interpreted as the proportion of variance explained by the model. Because MCA is based on the grand mean, all of these coefficients are derived from comparisons (ratios) of explained and unexplained sums of squares. For anyone familiar with analysis of variance, MCA is an easily tractable extension. For those not familiar with analysis of variance, the concept of variation around the means of categories offers an easily understood model building procedure.

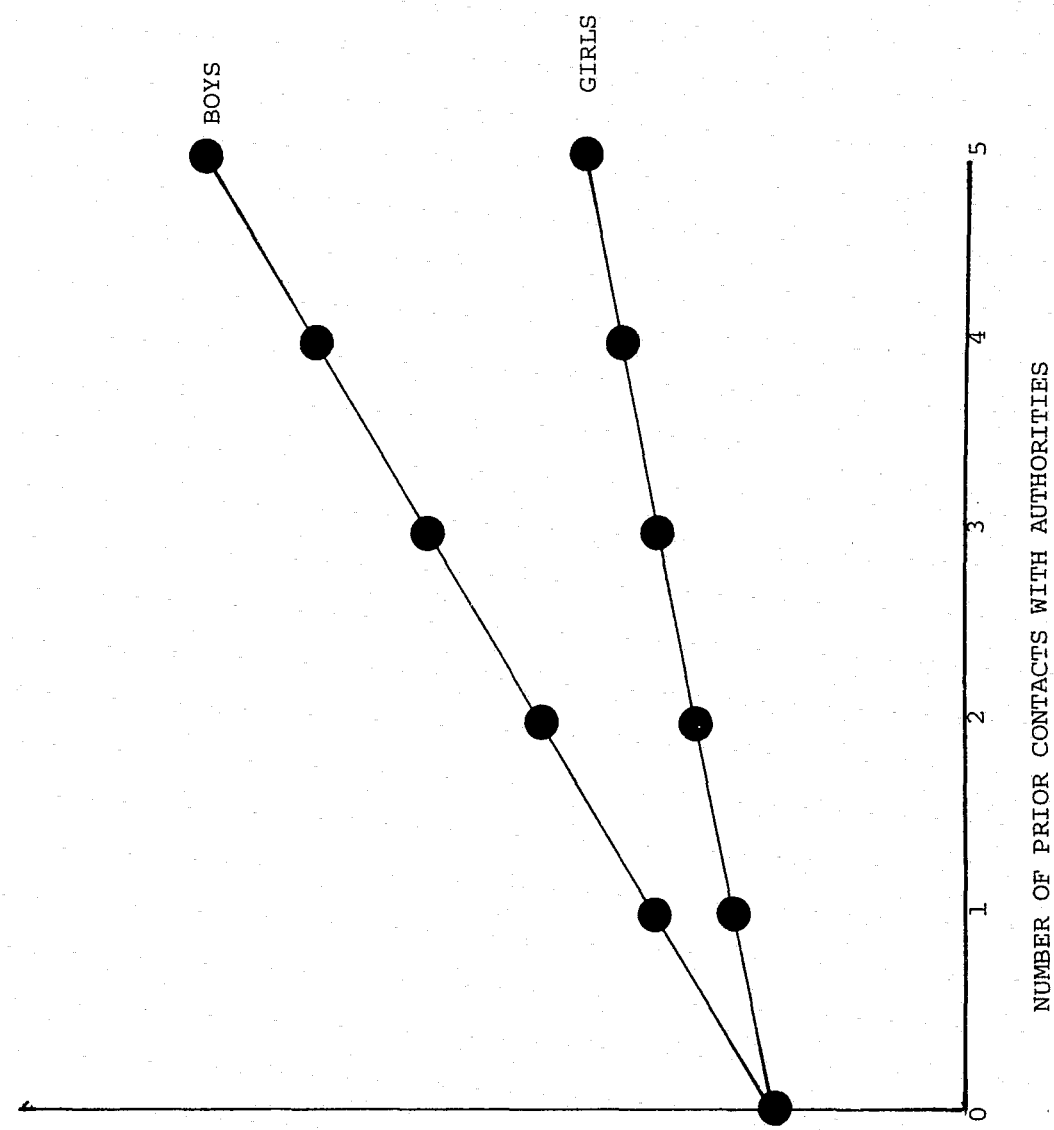
The Detection of Interaction Terms

The major weakness of MCA is its inability ot handle interaction terms. Moreover, MCA offers few clues that the additivity assumption has been violated. Therefore, it is possible to be misled by multiple classification procedures if they are applied blindly to a data set. This means that a preliminary analysis, searching for evidence of interaction terms, is generally required. This process has been automated by an elaborate computer routine known as Automatic Interaction Detection (AID) which is both powerful and complex. However, for a relatively small number of variables (on the order of four or five) much of the information gained from AID can be approximated by simpler and often more available procedures.

In simple terms, an interaction effect means that the variable A has one relation with variable Y for those subjects in the first category of variable B (b_1), but a different relation for those subjects in the second category of variable B (b_2). Thus we say variable B interacts with variable A to produce y. In the case of predicting recidivism, we might consider the relation of prior contact with authorities to recidivism for girls and boys. Assume for the moment that prior contact has a positive relation with recidivism for boys, as shown in Figure 3. Prior contacts also have a positive effect for girls, but the slope for girls is less than that for boys. Because the slope is different for the two categories of the sex variable, we say sex interacts with prior contacts in producing recidivism. A visual portrayal as in Figure 3 is probably the simplest device with which to check for interaction effects. It is also possible that an interaction effect may involve

FIGURE 3

HYPOTHETICAL RELATION SHOWING AN ORDINAL INTERACTION EFFECT



different signs or directions of slopes. In Figure 4 the relation between seriousness of the instant offense and recidivism is positive for boys but negative for girls. Because MCA does not impose linearity it is possible that some variables may take non-linear forms. However, it is still necessary to search for interaction terms. Figure 5 indicates the general principle involved--essentially we are looking for parallel and non-parallel lines. In Figure 5 the relation for boys is a steep inverted U-shape. The line for girls is not parallel, indicating a possible interaction effect involving a non-linear relation. The importance of interaction effects as well as non-linearity will vary with the data set being examined. If interactions are absent the next step is to proceed directly with multiple classification analysis.

If there is evidence of interactions, they must be included in the MCA model. This is usually accomplished by forming the cross product of the scores for each individual on the two interacting variables. This creates a new composite variable referred to as an interaction term. When an interaction term is included, the two variables from which it is created are usually excluded from the MCA model. For example, we might create an interaction term for sex and race. The values of the new interaction variable are shown in the table below.

		SEX	
		female	male
RACE	white	1	2
	non-white	3	4

FIGURE 4

HYPOTHETICAL RELATION SHOWING DISORDINAL INTERACTION EFFECT

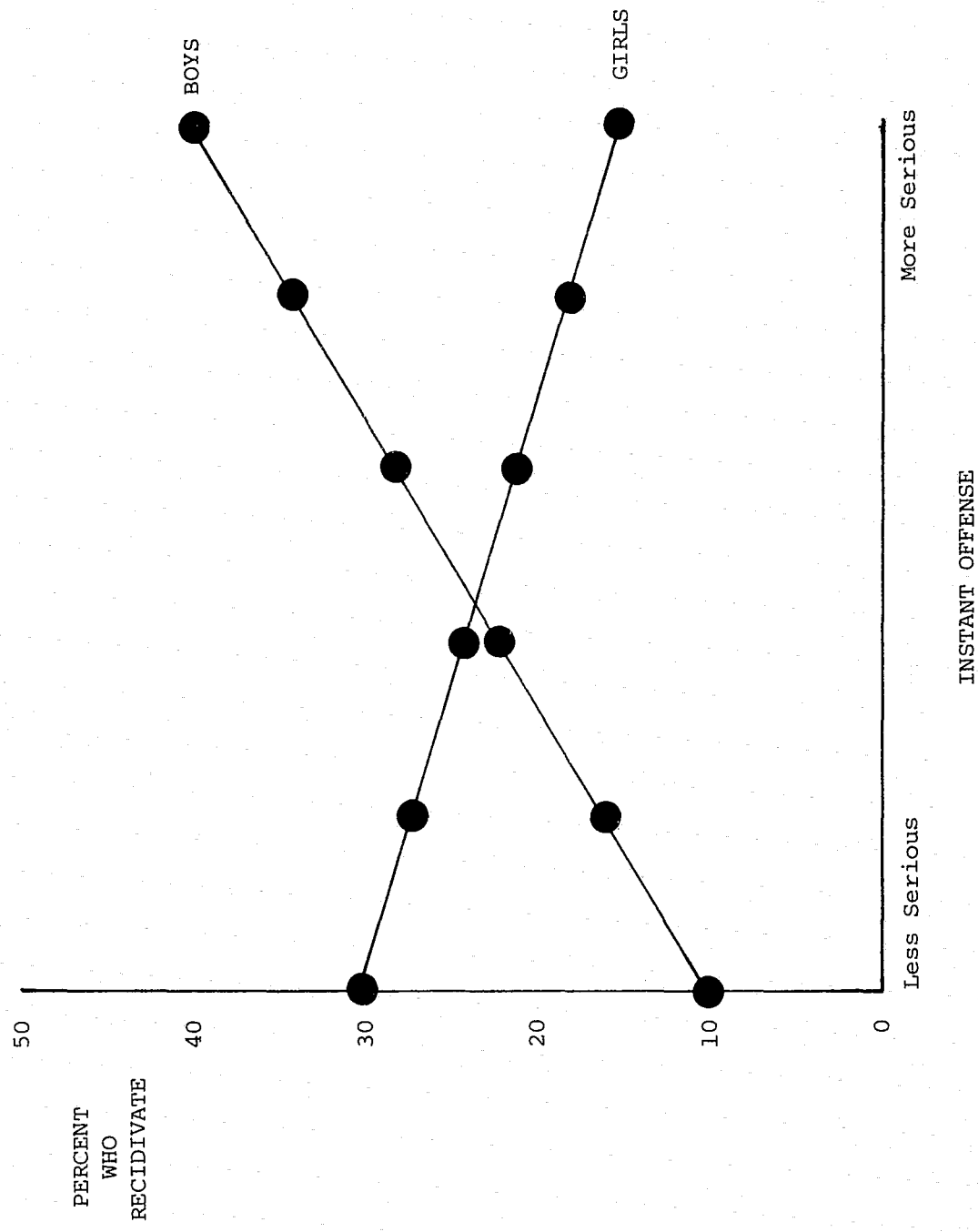
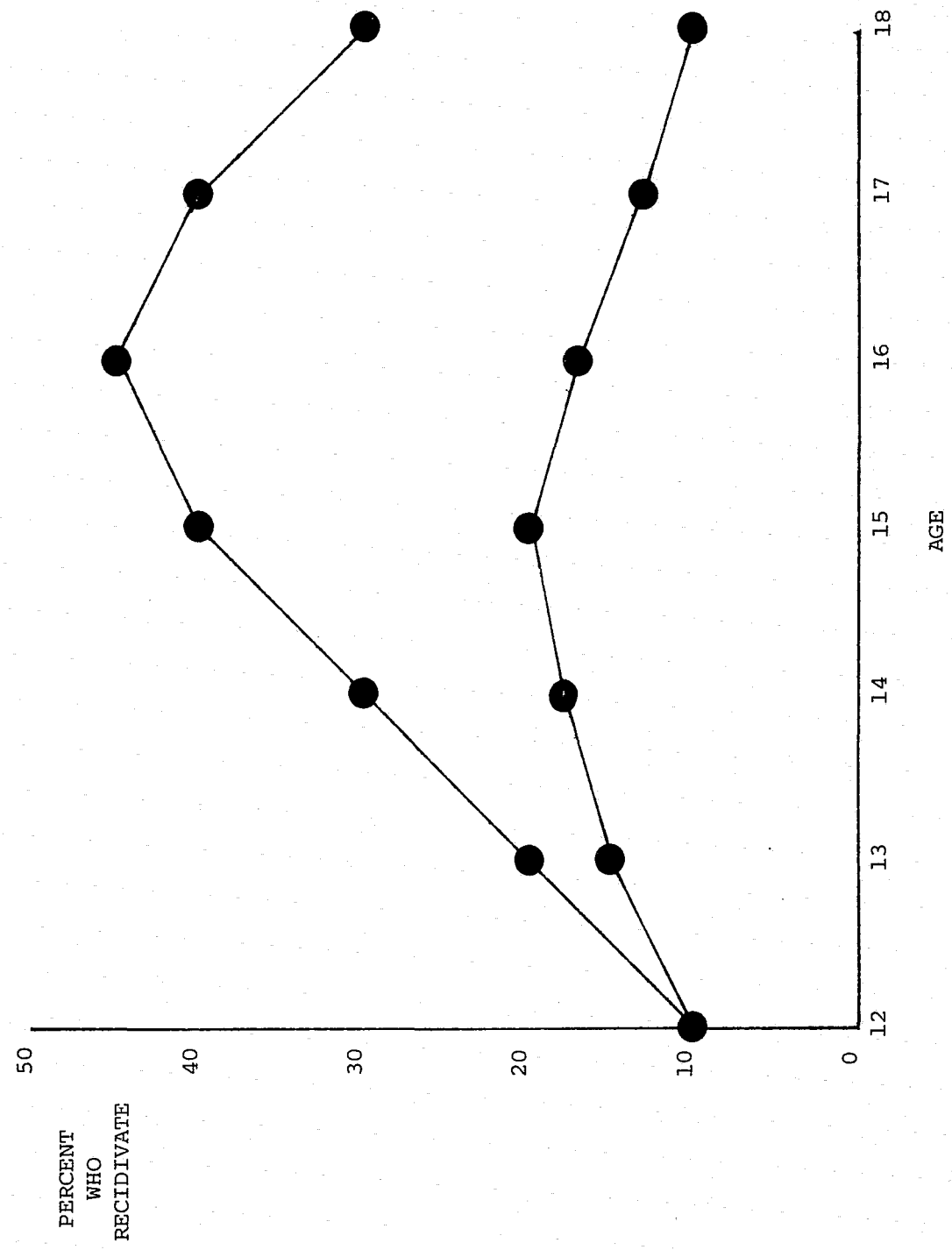


FIGURE 5

HYPOTHETICAL RELATION SHOWING AN INTERACTION EFFECT IN CURVILINEAR FORMS



The new variable has four categories and is entered into the MCA analysis directly. The results of the MCA analysis will estimate the predictive power of this "new" variable. It is possible that the new variable may be awkward or impossible to interpret. For example, there is no logical order among the four categories of the new variable. In fact, any set of numbers could have been given to the cells of the new variable. For prediction purposes it may be useful to assign cell values according to the order of category means on the criterion variable. Some interaction terms may make intuitive sense. For example, if we found evidence of an interaction between sex and race, we might simply consider this as four distinct types of youth culture: one for white girls, one for white boys, one for non-white girls, and one for non-white boys. In any event, once interaction effects have been detected and included, the use of MCA follows directly.

The process of interaction detection can be tedious with large numbers of predictor variables, which is why the AID program was originally developed. However, if there are only a few predictors it is possible to examine the pairs of predictors on a one-by-one basis (in general there will be $\frac{n(n-1)}{2}$ pairs of possible two-variable interactions where n is the number of variables). This can be accomplished by making use of the BREAKDOWN routine in SPSS. MCA can also be accomplished with SPSS by using the MCA option for the ANOVA routine.

Linear Regression Analysis

An alternative approach to constructing a predictive model is to use the simpler technique of linear regression. The general model can

be stated as: $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3...$ In this model it is assumed that the relation between each predictor variable and the criterion is linear. Regression directly averages all the categories of a given variable and creates a summary weight (the b coefficients) expressing the relative strength of each variable. Although these regression coefficients are partial coefficients which take into account the interrelations between the predictors, the model provides no evidence concerning the actual shape of the relation in the data. Like MCA, this model assumes a priori that the effects of the predictors are additive. Thus, the b weights and the intercept coefficient (a) are fitted to minimize the mean squared error (MSE) under the dual assumptions of additivity and linearity. The goodness of fit is expressed by the squared coefficient of multiple regression: $R^2 = 1 - MSE/Var_y$ where Var_y is the variance of the criterion variable.

Recent research indicates this simpler form of the multivariate function can produce robust predictions even though the assumption of linearity is violated. In general, it has been found that any monotone function can be well approximated by a line. For example, a power function such as $Y = X^2$ (shown in Figure 1) has a correlation of .975 with X (for the positive values of X). If a relation is clearly not a monotone (as shown in Figure 5), then the values of X can be rescaled in terms of their distance from the peak of the curve to create a new monotone function. Similarly, it has been found that for prediction purposes certain types of interactions can be predicted well by an additive function. Although various studies have reported "ordinal" interaction effects, it has been shown that multiplicative terms need not

be introduced into the regression model. As long as the slopes have the same sign, as shown in Figure 3, additive models simply average the two slopes and can be expected to produce excellent predictive results. Theoretically it is possible to encounter "disordinal" interactions or slopes with opposite signs, as shown in Figure 4. In practice this type of interaction is extremely rare and when encountered often is not replicable. However, if this interaction is encountered it needs to be transformed to a multiplicative term.

Use of an additive linear model when there is evidence of non-linear relations and/or interaction effects is contrary to intuition as well as the traditional logic of model building. However, parameter estimation from a single sample at a single point in time clearly capitalizes on chance variation in the sample. Any parameter estimated for a sample, such as the b weights of the regression equation, are biased toward the extreme values of the sample. Stated conversely, chance variation in the sample can produce parameters that are not at all indicative of the true parameters of the population. Thus, to create a non-linear and/or multiplicative model may not be warranted if these apparent deviations from the additive linear form are based on chance variation in the sample. Skepticism of sample estimates has led Dawes to argue that unless there is convincing evidence to the contrary, prediction models should be additive and linear. Dawes further suggests the relative weights of predictor variables should be equal. Thus, the major decision about the prediction model in this scheme is to decide if a variable should be included in the additive linear model.

The robustness of additive linear regression makes a powerful

argument for its use. It is both easy to apply and easy to interpret. Given these advantages and the excellent predictive results indicated by recent research, it is recommended that evaluators in need of a prediction for the untreated behavior of program populations consider this approach. The validity of the additive linear model rests on its ability to reasonably approximate conditionally monotone relations and ordinal interactions. This means that non-monotonic relations and disordinal interactions must be removed or transformed before the regression analysis is undertaken. As a multi-stage strategy it is suggested that the data display capacities of multiple classification analysis and then interaction detection can be employed as a first step to assure the conditions of conditional monotonicity and ordinal interaction have been met. Once these conditions have been assured, the final prediction can then be generated with the regression model.

FOOTNOTES

1. See annotated bibliography for basic discussion and documentation of multiple classification analysis computer programs.

2. This simplification is suggested only under the condition that the predictor variables all have positive intercorrelates. See article by Dawes and Corrigan (in annotated bibliography) for a complete discussion.

BIBLIOGRAPHY

The following citations are items on which the paper is based. This must be considered as a basic rather than exhaustive bibliography on the topic, as the literature on multivariate model building is extensive.

Interaction Detection and Multiple Classification Analysis

Multiple classification analysis is an extension of analysis of variance often referred to as analysis of covariance. The basic model has been recognized for some time, but computational aspects of extensive analyses were greatly simplified in the last fifteen years. Much of the early work in programming these procedures was done at the University of Michigan by Sonquist.

Andrews, Frank, James Morgan, and John Sonquist. Multiple Classification Analysis: A Report on a Computer Program for Multiple Regression Using Categorical Predictors. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, 1967.

Sonquist, John A. Multivariate Model Building: The Validation of a Search Strategy. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, 1970.

Sonquist, John A. and James N. Morgan. The Detection of Interaction

Effects: A Report on a Computer Program for the Selection of Optimal Combinations of Explanatory Variables. Monograph No. 35, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, 1964.

Linear Regression Analysis

Because linear regression has been one of the mainstays of social analysis for so many years, it would be impossible to present even a representation of the literature on the subject here. For a very basic presentation of multiple regression there are any number of texts to which the reader might turn depending on mathematical ability. For a rigorous but standard introduction, see:

Johnston, J. Econometric Methods, second edition. New York: McGraw-Hill, 1972.

The recent research referred to in this paper is reported in:

Dawes, Robyn M. and B. Corrigan. "Linear Models in Decision Making." Psychological Bulletin, 81, 1974, 95-106.

Einhorn, H.J. and R.M. Hogarth. "Unit Weighting Schemes for Decision Making." Organizational Behavior and Human Performance, 13, 1975, 171-192.

Wainer, H. "Estimating Coefficients in Linear Models: It Don't Make No Never Mind." Psychological Bulletin, 83, 1976, 213-217.

For a basic discussion of the computational aspects of both multiple classification analysis and multiple regression analysis, the reader is referred to:

Nei, Norman H. et al. Statistical Package for the Social Sciences, second edition. New York: McGraw-Hill, 1975.

SECTION 2F

APPLICATION OF ARIMA AND ANCOVA TO INTERRUPTED TIME SERIES*

Abstract

The major differences and similarities between the ARIMA and ANCOVA approaches to time series analysis are discussed in this paper. ARIMA is an acronym for auto regressive integrated moving averages and ANCOVA refers to analysis of covariance. The latter is used to demonstrate the rationale which underlies all of the deterministic approaches (linear trend predictions) to time series.

---

*This is a draft working paper written by Anne L. Schneider. Please do not quote from this draft until revisions have been made.

APPLICATION OF ARIMA AND ANCOVA TO INTERRUPTED TIME SERIES

Introduction

The purposes of interrupted time series designs are (a) to compare the slope (trend) of the pre intervention data with the slope (trend) in the post intervention data and ascertain whether a statistically significant change has occurred; (b) to compare the level (either the mean or the intercept value) of the pre with the post and test to determine whether a statistically significant change has occurred; and (c) to compare the entire pre intervention data with the post and test to determine whether the data in the post intervention time period are from a different population than the pre intervention data. The latter test incorporates both the level and slope in the test, whereas the first one seeks to compare slopes (holding level constant) and the second attempts to compare levels holding the slope constant.

The basic procedure is to predict or forecast from the pre intervention observations of Y (the dependent variable) into the post time period and then compare the predicted observations with those which actually are observed. In order to make these predictions or forecasts, one needs to develop a mathematical model for the pre intervention data which will make the most accurate possible forecasts. Thus, if the intervention has no effect on level or slope of the series, the predictions will correspond almost exactly to the observed values. We would conclude, then, that the intervention had no effect on the phenomenon being studied. But if the predicted values differ significantly from the observed values, then we could conclude that the differences are attributable to the intervention

(presuming that there are no other confounding effects).

Time series analysis normally involves four steps:

1. A preliminary identification of the mathematical model underlying the observations;

2. Estimation of the unknown parameters;

3. Diagnostic tests to determine if the model identified in step one is appropriate; and

4. Testing for statistical significance of the parameters, or forecasting, or using the model in whatever way it was intended.

Interrupted time series analysis should proceed with the same four steps, but the techniques are not nearly as well developed as they are for other types of time series analysis.

There are, in fact, a rather confusing plethora of methods and statistics for interrupted time series designs, including the Walker-Lev tests, analysis of covariance, the ordinary least squares (OLS) regression approach, the Chow test of significance when using OLS, and what has come to be called the Box-Jenkins approach that involves a series of different models known as ARIMA (p,d,q).[1]

The purpose of this paper is to present the four steps in normal time series analysis and, for each, to examine how the various approaches to interrupted time series designs has dealt with it.

Before proceeding, however, some equivalencies should be noted in the different approaches to interrupted time series designs.

1. Analysis of Covariance Approaches: In the ANCOVA approach, the pre intervention data are treated as group 1, the post data are group 2, and the covariate X is time (scored as 1, 2, 3, and so on). The Walker-Lev tests of significance for interrupted time series are identical to the usual ANCOVA tests. Furthermore, the ANCOVA procedures are identical

to ordinary least squares (OLS) regression analysis if time is used as the independent variable, the pre-post time periods are included as a dummy variable, and the interaction between time and the dummy variable is in the equation. The Chow test for significant differences of the regression coefficient is slightly different from the ANCOVA tests, but the differences are not substantial. Thus, in the subsequent discussion, all of these will be grouped together and discussed as the general ANCOVA approach to interrupted time series.

2. ARIMA (p,d,q): ARIMA is an acronym for autoregressive integrated moving average. Box and Jenkins first popularized these models and Glass, Willson, and Gotman first adapted them to interrupted time series analysis.[2]

Identification of the Mathematical Model

The first major problem in interrupted time series analysis is to identify the model underlying the data in the pre and post time periods. A key distinction between the ANCOVA approach and ARIMA is that the former assumes a deterministic model, whereas the latter assumes a stochastic structure but contains adaptations that can be used when the data follow a mixed stochastic/deterministic pattern. In the subsequent discussion, the two major types of deterministic models will be explained, as will the two major kinds of stochastic models. Following this is a presentation of how the ARIMA approach incorporates a deterministic element into the stochastic model and whether the ANCOVA approach could incorporate a stochastic element into the deterministic model.

Deterministic Models

The two most commonly found kinds of deterministic models are the constant mean model and the linear trend model. The form of the constant mean model is:

(1) $Y_t = \overline{Y}$ where Y = the original data, observed over several time periods

t = time, measured in months, years, etc.

$\overline{Y}$ = the mean of the Y values

he assumption of this model is that the post intervention data will be dequately predicted by the mean of the pre intervention data if the ntervention has no effect. Several authors (Box and Jenkins, Nelson) o not consider the constant mean model to be "deterministic." Box and enkins define a deterministic model as one in which the phenomenon is a function of time.

Linear trend is the second type of deterministic model. The form is:

(2) $Y_t = a_1 + b_1 t$ where Y = the observations

a_1 = the intercept value

b_1 = the regression coefficient

t = time, measured in months, years, etc.

This equation is solved with ordinary least squares (OLS) by regressing the actual values of Y on time. Time can be scored in any number of ways. One can use the year (e.g., time 1 might be 1965, time 2 1966, and so on) or one could number the time points, usually starting with "1" for the most distant and continuing through to the most recent time point. An equation of this type would be used for both the pre and post interven-

tion data, provided that there are sufficient data points to estimate the parameters a and b. (The OLS method of solving for a and b are the same as those used in analysis of covariance. For the ANCOVA procedure, the pre intervention data are treated as group 1, the post are group 2, and time is the covariate X.)

Non-linear trend models could be developed by transforming the original data in various ways, but the general technique is the same as described above.

One should notice that the equations have no error term, which, in practice, is unrealistic. Data could be considered to follow a generally deterministic model which leaves some error between the actual and predicted Y values, but the deterministic model should produce error that has the following two characteristics:

(a) The mean of the error is zero; and

(b) The error at one time point is not correlated with error at another time point.

To test this, one regresses the error at t with the error at t-1 for all successive pairs. This is shown below.

(3) $e_t = pe_{t-1} + v$ where p = the autoregression coefficient
e = error from prior equation
v = new error

The coefficient p is the autocorrelation coefficient for the residuals of the original equation and v is the new error term. If p is not significantly different than zero, one can assume that the errors are not autocorrelated and the deterministic model is an appropriate one for describing the data. (The Durbin-Watson test is used.)

If there is autocorrelation in the residuals (i.e., $p \neq 0$, within sampling error), then the appropriate model has not been identified. The consequences of this are very serious, because the equation not only produces an inferior forecast for the post time period, but the standard error of the regression line is seriously underestimated. In turn, this produces inflated F or t values when tests of significance are made for the regression coefficient.

The ANCOVA approach, then, assumes that a deterministic model underlies the data. The predictions are made from $Y = a + bt$. If the regression coefficient $b = 0$, the prediction of Y is equal to the intercept value which (when $b \neq 0$) is equal to the mean of Y.

In interrupted time series, the formula is:

$$(4) \quad Y = a + b_1 I + b_2 \text{TIME} + b_3 I\ \text{TIME}$$

where Y = the dependent variable

I = a dummy variable representing the intervention point; observations before the intervention would be given a score of zero; observations after the intervention would be given a score of one.

TIME = time, measured 1, 2, 3, 4, and so on to the most recent point, using weeks, quarters, years, etc.

I TIME = interaction between time and the intervention dummy variable (this variable is created by multiplying the score on the intervention variable [zero or one] and the score on the time variable, thereby creating a new variable for each case).

Stochastic Models

The second broad type of time series models, stochastic models, are considered to be appropriate when the phenomenon being studied is generally

random rather than being "determined" by trend or, for that matter, even the mean of the data. The observations of Y, if they follow a stochastic pattern, contain no trend and are not correlated with any other exogenous variable that has been measured and included in the model.

One way of conceptualizing this is that the phenomenon being studied is the product of random shocks which, once having occurred, influence the current value of the dependent variable (Y) and continue to have an effect (although a declining one) on future values of Y. Thus, the value of Y at one time point would be correlated with the Y values at one or more previous time points. The ARIMA approach is designed specifically for stochastic processes. There are two kinds of stochastic models, autoregressive models and moving average models.

Autoregressive Models (p)

The form of this model is:

$$(5) \quad Y_t - L = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + e_t$$

where ϕ = a parameter to be estimated

L = the mean or initial level of the series

The L values are deviations from the mean or from the initial level of the series (L). The symbol ϕ is the coefficient that governs the relationship between the value of Y at different time lags. Thus, ϕ_1 is the coefficient showing the relationship between pairs of Y values when Y at t is paired with Y at t-1 for all pairs. ϕ_2 is a partial coefficient showing the relationship between Y_t with Y_{t-2} when Y_{t-1} has been statistically controlled.

If a random shock has an impact only on the current value of Y and on the next one, then the second value of Y would be related to the first, but the third value of Y would not be related to the first. This would be an autoregressive scheme of order 1 (AR1). If a random shock has an impact that is felt for two time periods in the future, the model would be an autoregressive model of order 2 (AR2), and so on.

The coefficient $\emptyset$, if estimated using ordinary least squares, is the regression coefficient (b). There are, however, problems with using OLS to estimate coefficients when the independent variable is a lagged value of the dependent variable.[3] The major problem is that one cannot properly estimate $\emptyset$ until the error term contains no autocorrelation. But the Durbin-Watson test for determining whether autocorrelation of the error is significantly different than zero is not valid when a lagged value of the dependent variable is used as a predictor in the first equation. The preferred method is to estimate $\emptyset$ using maximum likelihood estimates (MLE) rather than ordinary least squares regression. The MLE procedures are used with the ARIMA approach. Conceptually, one takes all possible values of $\emptyset$ between -1 and +1 and tests each in the equation. The value of $\emptyset$ that minimizes the squared error is chosen.

Moving Average Models (q)

The form of the moving average model is:

$$(6) \quad Y_t - L = -\theta_1 e_{t-1} - \theta_2 e_{t-2} \ldots - \theta_q e_{t-q} + e_t$$

where θ = a parameter to be estimated

e = the error in the prediction at previous time points

L = the mean or initial level of the series

Notice that the difference between autoregressive (AR) models and moving average models (MA) is that the AR equation contains Y_{t-i} as an independent variable, whereas the moving average method contains the error in the previous prediction as the independent variable. By converting e to $Y_t - \hat{Y}$ (actual Y minus the predicted Y), the following formulae show the difference between the AR1 and the MA1 models.

(7) AR1: $Y_t - L = \emptyset_1 Y_{t-1} + e_t$

(8) MA1: $Y_t - L = -\theta_1 (Y_{t-1} - \hat{Y}_{t-1}) + e$

It obviously is difficult to estimate the coefficient θ because one cannot obtain a predicted value for Y at t-1 without knowing the value of θ. Yet, estimating the value of θ requires an estimate of the error term which is the difference between Y and predicted Y. This unending cycle is resolved with maximum likelihood procedures in which all possible values of θ_1 between -1 and +1 are tested in the MA1 model. Whichever value yields the lowest mean squared error (e_i^2) is selected.

The type of ARIMA model is indicated by the subscripts p,d,q. The first of these, p, refers to whether there is an autoregressive component in the Y values after removing the level or mean; d refers to whether the data were transformed using first differences, second differences, etc. (this will be discussed later); and q refers to whether there is a moving average component in the Y values. Thus, ARIMA (1,0,1) refers to a model in which p=1 (there is a first order autoregressive component); d=0 (first differences were not taken); and q=1 (there is a first order moving average component). A first order autoregressive component means that $\emptyset_1$ is calculated but $\emptyset_2$ is not. θ_1 is calculated but θ_2 is not in a first order

moving average model. These points are summarized below:

ARIMA (p,d,q)

p = the autoregressive component

d = the number of times differencing was used. d=1 means first differences, d-2 means second differences also were taken, and so on.

q = the moving average component

ARIMA (0,1,1) = first differences were taken on the original Y values and the predictions (forecasts) or underlying model is specified as $Y_t - L = -\theta_1 e_{t-1} + e_t$ with Y_t representing the values after first differences were taken.

ARIMA (1,0,1) = the predictions (forecasts) or underlying model is specified as having an autoregressive component of order 1. Using the original raw data, predictions are made in accordance with $Y_t - L = \phi_1 Y_{t-1} - \theta_1 e_{t-1} + e_t$.

Tests are made to determine that the new error (e_t) is not autocorrelated.

One basic distinction between ANCOVA and ARIMA approaches is that ANCOVA assumes a deterministic model in which Y is a function of time and the mean of the Y values, whereas ARIMA (as presented to this point) assumes that Y is a function of previous values of Y and/or previous errors in predictions of Y.

A second and very important distinction is that ANCOVA uses ordinary least squares regression, whereas ARIMA uses maximum likelihood techniques. (Provided that the model is a deterministic one, the OLS approach is appropriate; provided that the model is stochastic, the MLE approach is appropriate.)

Both approaches use the same diagnostic techniques for determining whether the model is appropriate. If the error term from the equation is not autocorrelated (that is, e_t is not functionally related to e_{t-1} or

to any other time lag), then the model is appropriate and one can proceed to test for intervention effects.

Differences between the two approaches become more complex and confusing if the underlying model is mixed and contains some deterministic elements as well as some stochastic ones.

Mixed Models

In general, the ARIMA approaches incorporate deterministic elements (if they exist) by:

(a) subtracting the mean or initial level (L) of the series from the Y values prior to calculation of $\emptyset$ or θ; and/or

(b) taking first differences in the Y values (or second differences) prior to calculation of $\emptyset$ or θ; and/or

(c) incorporating a constant term δ into the equation; it represents "drift" in the data through time.

It is fair to say that, in general, ANCOVA has no built in mechanism for incorporating stochastic processes into the model, but techniques for doing so using ordinary least squares have been suggested as an alternative (albeit an inferior one) to the use of maximum likelihood estimates used in ARIMA.[4]

For illustrative purposes, it will be useful to show the conditions under which ARIMA procedures for incorporating deterministic elements into the equation are similar to those used in ANCOVA. If we assume that there is no autocorrelation in the data (i.e., $\emptyset$ and θ are zero) and if we assume that the Y values are not a function of time (and contain no "drift"), then the two approaches are identical, as shown below:

(9) ANCOVA: $Y_t = a + bt + e$ $\quad b=0;\ a=\bar{Y}$

$Y_t = \bar{Y} + e$

(10) ARIMA: $Y_t - L = \emptyset Y_{t-1} - \theta e_{t-1} + \delta + e_t$ $\quad \emptyset=\theta=\delta=0;\ L=\bar{Y}$

$Y_t = \bar{Y} + e$

When the regression coefficient b = 0, alpha takes on the value of the mean of the series. Thus, the ANCOVA prediction is based entirely on the mean of Y. If $\emptyset$ and θ are zero and there is no drift in the data, the level of the series (L) is the mean of the data.

If we assume that there is drift in the data (which is analogous to short term trend) and if $\emptyset_1=1.0$ but $\theta=0$, then the two approaches are identical. Consider the following equations in which a=L=e=0, $\emptyset$=1.0, and t is measured in units of 1, 2, 3, etc:

(11) ANCOVA

$Y_t = a + bt$

$Y_t = bt$

$Y_3 = b3$

(12) ARIMA

$Y_t = L + \emptyset Y_{t-1} + \delta$

$Y_t = \emptyset Y_{t-1} + \delta$

$Y_3 = Y_2 + \delta$

$Y_2 = Y_1 + \delta$

$Y_1 = \delta$

Substituting:

$Y_3 = \delta+\delta+\delta$ AND $Y_3 = 3\delta$

Thus: $b = \delta$

It is important to notice that (in ARIMA) when a lagged value of the dependent variable is included as an independent variable, a trend or drift component is represented as a constant (to be estimated) rather than

as a parameter to be estimated and multiplied by time. The reason is that each previous value of Y already contains an appropriate "amount" of the constant, whereas in equations without lagged values the constant must be multiplied by the time variable. It also is important to notice that unless $\emptyset$ is exactly equal to 1.0 the inclusion of a "drift" variable in ARIMA models will not yield the same results as inclusion of a trend component (bt) in ANCOVA.

Apparently those employing the ARIMA models generally handle time dependency by taking first differences (or second differences, if needed) and selecting either the AR or MA model on the basis of a diagnosis concerning the autocorreation after differences have been taken on the original Y values. It can be shown that taking first differences is identical to a first order autoregressive model in which $\emptyset$=1.0.

This is shown below by comparing an ARIMA (1,0,1) model with $\emptyset$=1.0 to an ARIMA (0,1,1) model.

(13) ARIMA (1,0,1)

$$Y_t = \emptyset_1 Y_{t-1} - \theta e_{t-1} + e_t$$

OR

$$Y_t - \emptyset_1 Y_{t-1} = -\theta e_{t-1} + e_t$$

IF $\emptyset_1$=1.0, THEN

$$Y_t - Y_{t-1} = -\theta e_{t-1} + e_t$$

(14) ARIMA (0,1,1)

(first difference, moving average)

$$Y_t - Y_{t-1} = -\theta e_{t-1} + e_t$$

The technique of taking first differences, therefore, can be viewed as a method of incorporating time dependent (deterministic) elements into the model, but it actually is a special case of the autoregressive model.

Returning to the question posed earlier, it seems that the ARIMA models

are identical to the ANCOVA approach only for the ARIMA (0,1,0) model (first differencing within the equation) or a model that includes δ (drift) along with $\emptyset$=1.0 and θ=0. In addition, as noted earlier, ARIMA and ANCOVA are the same when there is no trend or drift in the data and no autocorrelation. In this case, both base the predicted Y values on the mean of the Y data.

An analogous question concerns what the investigator using ANCOVA should do if the residuals from the original OLS equation contain autocorrelation. The most commonly recommended procedure is to take first differences. Some authors, however, argue that first differences should not be taken unless the autocorrelation coefficient p (or $\emptyset$) actually is close to 1.0.[5] Otherwise, this procedure can result in erroneous conclusions. First differencing will remove a linear trend from the data (whether it actually was there or not). For example, if the values of Y follow a perfectly linear trend going from 10 at time 1 to 20 at time 2 and 30 at time 3, the first differences will be perfectly stationary (10, 10, 10). If the trend is greater than this (an exponential trend) then first differencing will leave some trend in the data and it is likely that the autocorrelation of the error term will still be significant.

As a substitute for first differences there is a procedure called IV-Pseudo GLS, which can be done with two OLS regression analyses.[6] The recommended approach, however, is to use maximum likelihood estimates of p.

This part of the discussion can be summarized as follows:

1. ANCOVA assumes the Y values are a function of time. If the residuals are autocorrelated (p>0), then the model is not an appropriate

one. Obviously, if there are predictable patterns in the errors, one could improve the predictions of Y by making use of this information. Furthermore, tests of significance for b will contain inflated F values.

2. There are no simple solutions to the autocorrelation problem using OLS. First differences calculation may not be a good solution.

3. The ARIMA models are far superior in terms of the ability to incorporate into the predictions a properly calculated coefficient relating values at one time point with those at another and, when needed, a coefficient that maximizes information contained in the error term. But when first differences are used on the original data, this approach has the same problem as noted above: The removal of trend for the pre and post restricts our ability to test for significant changes in trend or drift that might be attributable to the intervention.[7] However, ARIMA models that incorporate an autoregressive component (rather than first differences) and a constant representing drift would bypass this problem.

All of the previous discussion focused on identifying the model, estimating parameters, and diagnosing the fit of the model.[8] The problem of how one tests for significant changes in the level or slope of the data is discussed in the next section.

Testing for Intervention Effects

The ANCOVA approach for testing the effect of the intervention will be explained first, followed by a presentation of the technique recommended by Glass, Willson and Gotman for testing significance when using ARIMA models.

ANCOVA Tests

It should be noted at the outset that the three tests developed by Walker-Lev for interrupted time series analysis are identical to the standard analysis of covariance methods for testing significant differences. These, in turn, are identical to OLS tests using a dummy variable procedure to represent the pre and post time periods. Table 1 shows the formulae for Walker-Lev 1, Walker-Lev 3, and the Chow test using a regression (rather than ANCOVA) notation scheme. The reader should study the definitions for the terms used in Table 1 (see Table 2).

The equation involving Y_T is simple to calculate using standard regression procedures (and all of the data, pre as well as post). Likewise, the Y_i equations are simple to calculate since one uses standard regression on the pre I data for Y_1 and on the post I data for Y_2. Use of the three variable equation described previously yields all the information needed.

The F test is used to establish the probability of differences between pre and post time periods. In general, the value of F is found by dividing the explained sum of squares by the unexplained sum of squares. The unexplained sum of squares (USS) is the squared error found by subtracting the predicted Y from the observed Y. The total sum of squares (TSS) is found by subtracting the mean of all the Y values from each observation, squaring, and summing (TSS = ESS + USS).

Walker-Lev Test 1 (which is the same as the first analysis of covariance test) is designed to determine whether the slope in the pre time period is different from the post. The numerator of the F ratio consists of the difference between the predicted Y values from the equation using

TABLE 1

F TESTS[1]

Walker-Lev 1	$F = \frac{\Sigma(\hat{Y}_w - \hat{Y}_i)^2}{\Sigma(Y - \hat{Y}_i)^2} \cdot \frac{N - 2K}{K}$	Test for difference in slopes.
Walker-Lev 3	$F = \frac{\Sigma(\hat{Y}_w - \hat{Y}_t)^2}{\Sigma(\hat{Y}_w - \hat{Y}_i)^2 + (Y - \hat{Y}_i)^2} \cdot \frac{N-K-1}{K-1}$	Test for difference in intercepts.
Chow Test	$F = \frac{\Sigma(\hat{Y}_T - \hat{Y}_i)^2}{\Sigma(Y - \hat{Y}_i)^2} \cdot \frac{N - 2K}{K}$	Test for difference in entire regression line (intercept and slope).

[1]The numerator for Walker-Lev 1 in expanded form is: $\Sigma\left[(\hat{Y}_w - \bar{Y}_i) - (\hat{Y}_i - \bar{Y}_i)\right]^2$

TABLE 2

REGRESSION NOTATIONS

(X = time)

1)	$Y_T = a_T + b_T X + e$	b_T = regression of Y on time for entire series. b_T and a are best coefficients for all the data.
2)	$Y_{i\ i=1} = a_i + b_i X + e$	$b_i = b_1$ = regression of pre-intervention Y values on pre-intervention time points (separate group regression).
3)	$Y_{i\ i=2} = a_i + b_i X + e$	$b_i = b_2$ = regression of post-intervention Y values on post time points (separate group regression).
4)	$Y_{w\ w=1} = a_1 + b_w X + e$	b_w = best <u>common</u> slope for both pre and post; a_1 = intercept for preintervention data.
5)	$Y_{w\ w=2} = a_2 + b_w X + e$	b_w = best <u>common</u> slope for both pre and post; a_2=intercept for post-intervention.
6)	$e_t = a + pe_{t-1}$	p = autoregression coefficient showing serial dependency in error from any of the regressions described above.
7)	$e_{T_t} = a_T + p_T e_{T_{t-1}}$	p_T = autoregression coefficient from Y equation (also called total error, slope and intercept removed).
8)	$e_{i_t} = a_i + p_i e_{i_{t-1}}$	p_i = autoregression coefficient for equations 1 and 2 (also called separate group, slope and intercept removed).

the best common slope (b_w) and the equation using the best slope for each time period calculated separately (b_i). Then the difference between predictions based on b_w from those based on b_i is attributable to differences in slopes. Therefore, if $\hat{Y}_w - \hat{Y}_i = 0$, the slopes are the same for both groups. If this is greater than zero, we simply assume that the slope based on the separate group regressions is more accurate than the common slope. In a sense, the Walker-Lev test 1 shows whether the "gain" in explained variation provided through the use of a unique slope for each time period is significantly different than zero. If so, we assume that the slopes are different. If not, we assume that the common slope b_w is an adequate description for both the pre and post data.

Walker-Lev 3 compares the predicted values based on a common slope ($\hat{Y}_w$) with those based on using one regression equation for the entire pre and post data ($\hat{Y}_t$). If the numerator shown in Table 1 for the Walker-Lev 3 test is zero, this indicates that there is no gain in explained variation from using a common slope (but unique intercepts) over using one intercept value and a slope estimated from all data points. If the slopes are the same in the pre and post time periods, then all differences between predictions from Y_w and from Y_T will be attributable to differences in the intercept. Thus, when the slopes are equal, Walker-Lev 3 tests for significant differences in alpha (or the level of the series between pre and post). When the slopes are not equal, the Walker-Lev 3 is not particularly meaningful. It appears, in fact, as if the denominator shown in Table 1 should not be used unless the slopes are equal.[9]

The Chow test involves a comparison of the regression line calculated from the entire set of data (pre and post) with the regression line

for each time period. The purpose is to determine if the post I data are from a different population than the pre I data. If the F test is significant, we do not know whether the differences between pre and post are attributable to differences in level (intercept) or to differences in slope. Nevertheless, the Chow test is a straightforward method of determining whether the intervention had a significant effect.

The explained sum of squares used in any of the equations will be over-estimated if there is autocorrelation in the residuals from the original regression of Y on time. Although this presents no particular problem in the numerator of the statistics, since the ESS from one regression line is subtracted from the ESS of another, the denominator containing the unexplained sum of squares is underestimated when ESS is overestimated. Thus, the Walker-Lev statistics and the Chow test for significance cannot be relied on when there is autocorrelation remaining in the residuals of the original regression equation.

Tests of Significance for ARIMA Models

Those who have adapted ARIMA models for use in interrupted time series have used t tests to determine whether the initial level of the series is greater than zero and to determine the significance of δ.[10]

As noted previously, incorporation of δ into the equation when an AR1 model is being used (and $\emptyset$=1.0) results in δ taking on a value analogous to the trend component in the ANCOVA models. However, when a moving average model is used, the value of δ does not cumulate over time and it becomes a measure of change in the level of the series. Thus, the current state of the art in using ARIMA models for interrupted time series results

in there being no test for change in trend, but only a test for change in the level of the series.

The problem of how to incorporate a test for change in trend (or drift) is difficult to resolve. One possibility would be to use an autoregressive component in the model whenever any drift or trend is apparent (rather than taking first differences) and to always include δ. If δ is not significantly different than zero, when previous values of Y are in the equation, then one could conclude that there is no incremental shift upward or downward from one time point to the next. If $\emptyset$ is equal to 1.0 or -1.0, however, this model is identical to one which contains a linear trend component (bt). If $\emptyset$ is not equal to $\pm$1.0 (it would fall between -1 and +1), then some of the drift apparently would be measured with $\emptyset$ and some of it with δ.

FOOTNOTES

1. For discussions of the ANCOVA approaches see Helen M. Walker and Joseph Lev, STATISTICAL INFERENCE (Holt, Rinehart, and Winston, 1953); Charles W. Ostrom, Jr., TIME-SERIES ANALYSIS: REGRESSION TECHNIQUES (Sage Publications, 1978); Joyce Sween and Donald T. Campbell, THE INTERRUPTED TIME SERIES AS QUASI-EXPERIMENT: THREE TESTS OF SIGNIFICANCE (Vogelbach Computing Center, Northwestern University, 1965); and William L. Hays, STATISTICS FOR PSYCHOLOGISTS (Holt, Rinehart, and Winston, 1963).

   The ARIMA models are discussed in George E.P. Box and Gwilym M. Jenkins, TIME SERIES ANALYSIS: FORECASTING AND CONTROL, revised edition (Holden-Day, 1976); Charles R. Nelson, APPLIED TIME SERIES ANALYSIS (Holden-Day, 1973); Warren Gilchrist, STATISTICAL FORECASTING (John Wiley & Sons, 1976); Gene V. Glass, Victor L. Willson, and John M. Gottman, DESIGN AND ANALYSIS OF TIME-SERIES EXPERIMENTS (Colorado Associated University Press, 1975); and Stuart J. Deutsch and Francis B. Alt, "The Effect of Massachusetts' Gun Control Law on Gun-Related Crimes in the City of Boston," in EVALUATION QUARTERLY, 1 (1977), 543-568.

2. Box and Jenkins, op. cit., and Glass et al., op. cit.

3. Ostrom, op. cit., has a good discussion of this.

4. See Ostrom, ibid., for a discussion of how this can be done using SPSS.

5. Ibid.

6. Ibid.

7. The interpretation of the types of change (change in trend or in level) is difficult when first differences have been taken.

8. There are methods of examining the lag correlations of the original (raw) data which are intended to identify the model. See Nelson, op. cit., or Gilchrist, op. cit.

9. See Hays, op. cit., for a discussion of ANCOVA.

10. See Glass and Deutsch, op. cit.

SECTION 2G

DETERMINING APPROPRIATE SAMPLE SIZES IN EVALUATION*

Abstract

Tables showing the sample size needed in order to achieve statistical significance at the .05 and .01 levels for the Z test of proportions are included. The determination depends on the magnitude of the proportion. When this is not known, the evaluator would need to estimate the expected percentage in order to estimate the size of sample needed. Similar tables are more difficult to construct or use for other types of significance tests because the variance and mean of the data have to be known or estimated.

*This paper was prepared by William R. Griffith and has been accepted for publication in Victimology.

DETERMINING APPROPRIATE SAMPLE SIZES IN EVALUATION

Evaluators frequently encounter the problem of determining the appropriate sample size for their research. Often, arbitrary and capricious criteria are employed when sample sizes are selected, resulting in samples which are either too small to enable the evaluator to detect the hypothesized treatment effect or so large that human and financial resources are wasted.

This paper presents two tables which will enable evaluators to determine more accurately the appropriate sample size prior to the conduct of their research. Specifically, these tables present the minimum number of cases needed for obtaining statistical significance between two proportions; Table 1 reports values for a .05 significance level and Table 2 is for a .01 significance level.

The following formula was used in generating the values in these tables:

$$N = \frac{z^2(p_1q_1 + p_2q_2)}{(p_1 - p_2)^2}$$

where: N = number of cases in both pre and post samples (independent samples)

z = Z value for a one-tailed test of statistical significance ($z = 1.645$ for $\alpha = .05$; $z = 2.325$ for $\alpha = .01$)

p_1 = proportion for the "pre" period (or group 1)

p_2 = proportion for the "post" period (or group 2)

$q_1 = 1 - p_1$

$q_2 = 1 - p_2$

One would use the tables in the following way: Let us say that one is evaluating the effectiveness of a burglary reduction program through the use of a victimization survey. The estimated burglary victimization rate prior to the implementation of the program is 9 per 100, and the goal of the program is to reduce the burglary rate by 11% (i.e., reduce the rate to 8 per 100). Thus, in this case, $p_1 = .09$ and $p_2 = .08$. For the differences between .09 and .08 to be statistically significant at the .05 level, Table 1 shows that a minimum of 4,208 valid interviews would be necessary for both the pre and post surveys; at the .01 level, Table 2 shows that 8,413 cases are necessary in each survey.

In addition to pre and post samples, these tables can be used for determining the sample size for any two groups. For example, if 25% of group A and 30% of group B expressed dissatisfaction with the job that their local law enforcement agency was doing, Table 1 shows that a minimum sample size of 430 for each group is necessary in order to obtain statistical significance ($\alpha = .05$) between these two proportions.

In a similar fashion, given a certain number of cases, one can use the tables to determine what differences in proportions would be statistically significant. For example, if one had pre and post samples composed of 1,000 cases each, one could anticipate that differences of 9% and 7% would be significant at the .05 level, while differences of 9% and 8% would not be significant. Thus, a program which allocates resources for a pre and post victimization survey of five hundred randomly selected respondents each and which anticipates that the burglary rate will drop from 9 per 100 households annually to 8 per 100 households annually would be unable to detect such a change given the proposed sample size.

As a demonstration of this formula, a test of significance ($\alpha = .05$) is calculated below for the two proportions and the two sample sizes given in the first example.

The test of significance between two proportions is computed by the following formula:

$$Z = \frac{\bar{\bar{X}}_1 - \bar{X}_2}{\sqrt{\bar{p}\ (1 - \bar{p})\ (1/N_1 + 1/N_2)}}$$

where: $\bar{p} = (N_1\bar{X}_1 + N_2\bar{X}_2)/(N_1 + N_2)$

N_1 = number of interviews in group 1

N_2 = number of interviews in group 2

$\bar{X}_1$ = group 1 proportion

$\bar{X}_2$ = group 2 proportion

For the example above, the values would be:

$N_1 = 4,208$

$N_2 = 4,208$

$\bar{X}_1 = .09$

$\bar{X}_2 = .08$

$\bar{p} = .085$

Thus,

$$Z = \frac{.09 - .08}{\sqrt{.085\ (.915)\ (1/4,208 + 1/4,208)}}$$

$$= \frac{.01}{\sqrt{.085\ (.915)\ (.0004753)}}$$

$$= \frac{.01}{.00608}$$

$Z = 1.645$ ($\alpha = .05$, one-tailed test)

It should be noted that these are the minimum number of cases necessary in each sample in order to obtain statistically significant differences between two proportions. In field research settings there will always be a number of "missing" cases which will vary depending on the types of questions being asked and the characteristics of the population being sampled; thus, such factors must be taken into account when using these tables. Moreover, for all proportions where the suggested number of cases is less than one hundred for each of the two samples, one is advised to increase the sample size by at least ten percent in order to compensate for discontinuities associated with small Ns.

TABLE 1

NUMBER OF CASES NEEDED FOR STATISTICAL SIGNIFICANCE BETWEEN TWO PROPORTIONS

(α = .05)

	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
.02	798																										
.03	264	1318																									
.04	145	392	1827																								
.05	97	202	518	2324																							
.06	72	129	257	641	2812																						
.07	56	92	159	311	762	3288																					
.08	46	70	111	189	364	879	3753																				
.09	39	56	83	130	219	416	994	4208																			
.10	33	46	66	97	149	248	466	1107	4652																		
.15	19	24	29	37	47	61	81	111	157	235																	
.20	13	15	18	21	25	30	36	44	54	68	311																
.25	-	11	12	14	16	18	21	24	28	33	85	376															
.30	-	-	-	10	11	13	14	16	18	20	41	100	430														
.35	-	-	-	-	-	-	10	11	12	14	24	47	112	474													
.40	-	-	-	-	-	-	-	-	-	10	16	27	51	122	506												
.45	-	-	-	-	-	-	-	-	-	-	11	18	29	55	129	528											
.50	-	-	-	-	-	-	-	-	-	-	-	12	19	31	57	133	538										
.55	-	-	-	-	-	-	-	-	-	-	-	-	13	20	32	59	134	538									
.60	-	-	-	-	-	-	-	-	-	-	-	-	-	14	20	32	59	133	528								
.65	-	-	-	-	-	-	-	-	-	-	-	-	-	10	14	20	32	57	129	506							
.70	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	14	20	31	55	122	474						
.75	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13	19	29	51	112	430					
.80	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	18	27	47	100	376				
.85	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	16	24	41	85	311			
.90	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	14	20	33	68	235		
.95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	16	25	47	149	

TABLE 2

NUMBER OF CASES NEEDED FOR STATISTICAL SIGNIFICANCE BETWEEN TWO PROPORTIONS

(α = .01)

	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
.02	1596																										
.03	527	2635																									
.04	290	784	3652																								
.05	194	403	1036	4647																							
.06	143	257	514	1282	5621																						
.07	113	183	319	622	1523	6573																					
.08	92	140	222	379	728	1758	7504																				
.09	78	112	167	260	438	831	1988	8413																			
.10	67	93	131	193	298	495	932	2213	9300																		
.15	38	47	59	74	95	123	163	222	315	471																	
.20	25	30	35	42	50	60	72	88	108	135	622																
.25	19	21	24	28	32	37	42	49	57	67	170	752															
.30	14	16	18	20	22	25	28	32	36	41	81	200	860														
.35	11	12	14	15	17	18	20	22	25	27	48	93	225	947													
.40	-	10	11	12	13	14	15	17	18	20	32	54	103	243	1012												
.45	-	-	-	-	10	11	12	13	14	15	23	35	59	110	257	1055											
.50	-	-	-	-	-	-	-	10	11	11	17	25	38	62	115	265	1077										
.55	-	-	-	-	-	-	-	-	-	-	13	18	26	40	64	117	268	1077									
.60	-	-	-	-	-	-	-	-	-	-	10	14	19	27	40	65	117	265	1055								
.65	-	-	-	-	-	-	-	-	-	-	-	10	14	19	27	40	64	115	257	1012							
.70	-	-	-	-	-	-	-	-	-	-	-	-	11	14	19	27	40	62	110	243	947						
.75	-	-	-	-	-	-	-	-	-	-	-	-	-	11	14	19	26	38	59	103	225	860					
.80	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	14	18	25	35	54	93	200	752				
.85	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	13	17	23	32	48	81	170	622			
.90	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	15	20	27	41	67	135	471		
.95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	13	17	22	32	50	95	298	

SECTION 2H

AN INTRODUCTION TO RELIABILITY & VALIDITY PROBLEMS IN CRIMINAL JUSTICE EVALUATION*

Abstract

Measurement error can be produced either by a lack of reliability or by a lack of validity in the data; much of the data used in criminal justice evaluation suffers from one or both problems. The impact of measurement error on the results depends on whether the error is correlated with values of the independent variable or whether it is randomly distributed vis a vis the independent variable. In the former situation, the error can distort or even reverse the true direction of the relationship. If the error is not correlated with the independent variable, the major consequence is that values of statistics such as F, t, Z, and the values of parameters such as the correlation coefficient, and others that are based on explained and unexplained variation are underestimated.

*These materials are a revision of those prepared by Anne L. Schneider and L.A. Wilson II which were originally presented at a special forum for ALJE evaluators.

CONTINUED

2 OF 7

AN INTRODUCTION TO RELIABILITY & VALIDITY PROBLEMS IN CRIMINAL JUSTICE EVALUATION

Introduction

Measurement problems are inherent in virtually all empirical social science research. Even though the terms reliability and validity generally are associated with basic research rather than evaluation research, they are equally relevant to both. Propositions about the effect that a new program will have upon a treatment population arise from some theory about the relationship which exists between the selected treatment and a particular behavior. It is in the translation of the concepts from the theory into observable behavior, events, or predispositions that the issues of reliability and validity arise.

The importance of minimizing error in the measurement of the concepts should not be underestimated. If the measures contain error that is correlated with values on the independent variables, then the likelihood is increased that the investigator will conclude that the treatment and control groups are significantly different when, in fact, they are not. (This is a Type 1 error: The false rejection of a true null hypothesis.) If the measures contain error that is not correlated with values on the independent variable, then the investigator may find no significant differences when, in fact, the differences were significant. (This is a Type 2 error: The failure to reject a false null hypothesis.)[1] While not all of the "nothing works" literature can be understood in these terms, it is reasonable to assume that some of the failure to find significant change as a result of program implementation is a

function of uncorrelated measurement error which depresses the value of tests of significance and, therefore, results in an unwarranted conclusion of "no effect."[2] These points will be illustrated in a subsequent section of the paper, following a presentation of what is meant by "reliability" and by "validity" of measurement.

Reliability and Validity in Criminal Justice Research

The concepts of reliability and validity are most often associated with attitude measurement. Different types of reliability (consistency and stability) and validity (content, predictive, construct, convergent-discriminant, etc.) have been identified and methodologies and statistical models developed to assess them.[3] The purpose of this paper is to extend the concern about reliability and validity to the behavior and event data more frequently dealt with in the evaluation of criminal justice programs.

Reliability has two different meanings. First, it can refer to the consistency or uni-dimensionality of a set of items used to measure some phenomenon. In attitude measurement it is frequently assumed that a fairly complex phenomenon is under investigation, such as alienation, and multiple items are required to operationalize the concept. Since each of the items is designed to measure the same concept (with slightly different aspects of the concept being dealt with by specific items), it is assumed that a reliable set of items will have a relatively high average inter-item correlation. Statistics such as the Kuder-Richardson formulas 20 and 21 and the Cronbach alpha are used

to assess this type of reliability.[4]

The second definition of reliability refers to the stability of the observations that are made. In this case, reliability is assumed to exist if, for instance, a respondent always gives the same answer to the same question, assuming that conditions have not changed which would explain a change in respondent reply. It is this latter interpretation of reliability that has the greatest relevance for event and behavior data.

Validity refers to whether or not one is measuring the concept that is presumably being measured. If there are clear behavioral referents in the definition of a concept, the assessment of its validity can be rather simple, such as correlating it with some other variable to which it should be related (predictive validity). For more abstract concepts that do not have clear behavioral referents, an assessment of validity can involve a demanding specification of a whole series of relationships that should be expected (construct validity).

The application of the terms validity and reliability to behavior or event data in criminal justice might be demonstrated by the problem of measuring the incidence of burglary in society. Addressing the problem of validity first, the measure of burglary (1) should accord with our best understanding of the concept and (2) should measure the same thing in each criminal justice system in the nation. As Przeworski and Teune note, "In a comparative or cross systemic context, validity means that we are measuring in <u>each system</u> under consideration what we intend to measure."[5]

The valid measurement of burglary would seem to be plagued with

two immediate problems, both of which have to do with the definition of the concept. First, as the International Association of Police Chiefs were aware in 1927, the thousands of police jurisdictions in the United States had hundreds, if not thousands, of idiosyncratic definitions of what constituted a burglary or other major crime. When this association's Committee on Uniform Crime Records published their manual in 1929, their explicit purpose was to bring uniformity to the definition of different types of crime. This was clearly the first step toward the development of valid measures of crime for society.

The second problem in the development of a valid measure of burglary has to do with our understanding of what is actually being measured. Assuming that all jurisdictions are judiciously abiding by the guidelines of the Uniform Crime Reporting Handbook, official police data reflect only the rates of reported crime, not all crime in society. Hence, such measures can only be valid if they are qualified by the term "reported."

Although the criteria used to identify and enumerate the incidents of reported burglary in society may be valid, substantial opportunities for unreliability in the actual figures are known to exist. The sources of this unreliability can be found in the carelessness of crime codes as well as in the purposeful misrepresentation of information. The latter source of unreliability, in fact, led the FBI to withhold publication of crime statistics during the years 1949 through 1952 for New York City--precincts in that city were grossly underreporting the incidence of all crime.

As will be discussed in the following section, both reliability

and validity refer to the existence of error--either systematic (correlated) or random (uncorrelated)--in our measurement. Either type of error can lead to the wrong inferences being made about the success or failure of a program being evaluated.

Effect of Low Validity or Reliability on Evaluation Results

The first way in which the lack of reliability or lack of validity can influence the results from evaluations is that it can introduce bias into the direction of the relationship. This can be illustrated with an example in which the concept the investigator wishes to measure is the total number of delinquent offenses committed by juveniles in the experimental and control groups. The actual indicator used is the number of re-contacts with the juvenile court. The variable--re-contact with the juvenile court--contains considerable error when it is used to measure the number of delinquent offenses actually committed, as shown in Table 1. Low reliability or validity can influence the direction of the relationship and confuse the interpretation as to which group had lower delinquency rates. In Table 1 the true proportion of the experimental and control groups committing subsequent delinquent offenses is 40 percent and 20 percent, respectively. Based on this measure of recidivism, the experimental treatment is not effective. But suppose that the police, for one reason or another, always refer any youth in the control group who committed a subsequent offense to the juvenile court, but only refer a fraction of the youths in the experimental group who commit subsequent offenses to the court. If this happened, then the observed re-contact measures could be reversed, as shown in Table 1, so that the

TABLE 1

HYPOTHETICAL DATA

ON OBSERVED AND TRUE RECIDIVISM

WITH CORRELATED ERROR[1]

GROUP	OBSERVED RECIDIVISM (Re-Contact with Court)	TRUE PROPORTION COMMITTING SUBSEQUENT OFFENSE
Experimental Group	10%	40%
Control Group	20%	20%

[1]In this example, the bias is introduced because the true measure was not used and because of a referral process (to the court) that was correlated with the treatment condition.

recidivism rate for the experimental group (10%) is lower than for the control (20%). Thus, the use of re-contact scores rather than true subsequent offenses produced a reversal in true recidivism differences between experimental and control groups.

Whether a reversal in the direction of the relationship of this kind is likely to occur depends on whether the lack of reliability or validity has the same effect on both groups. In the previous example, it did not. When the error does have the same effect on both groups the measurement problem can be treated as a special type of threat to validity. With a strong experimental design (and no treatment interaction effects of the type described in the previous example), one could be more assured that measurement problems did not alter the true direction of the relationship, but with weak designs one would not be as confident of this.

It also is important to insure that the methods of collecting and reporting data are the same for the experimental and control groups. If so, then even though there may be some unreliability in the measures one could be more confident that the error affects both groups in the same way.

If there are reasons to believe that the reliability and/or validity problems will not be the same for the experimental and control groups and will, therefore, distort the apparent differences between them, the evaluator should consider the following steps in order to strengthen the likelihood of being able to draw accurate conclusions:

1. If the problem stems from different data collection procedures or methods across the groups or areas, the evaluator should arrange for data to be collected with the same instruments (and, if possible,

the same people). At a minimum, the evaluator should ensure that the instruments are the same and, if there are different people collecting the data for the experimental and control groups or areas, the evaluator should train them to use the same techniques and should conduct a reliability check among them.

2. If the problems arise from a weak design and could be corrected with a stronger research design, then the evaluator should attempt to implement this solution before the project is so far underway that no changes can be made.

3. In case nothing can be done about the problem, the evaluator's responsibility is to assess the nature of the bias, measure it precisely (if possible), and adjust the conclusions accordingly. This involves, first, an assessment of whether the bias is such that it would make the project appear to be more effective than actually observed or less effective. If the results of the evaluation indicate the project was effective and the measurement problem is such that it would produce an underestimate of the true effectiveness, then the evaluator can conclude that the results are a conservative estimate of true project effect. On the other hand, if the results indicate the project is effective but the bias works in such a way as to make the project appear more effective than it actually is, the evaluator will not be able to draw any conclusions about project effectiveness.

The second way in which measurement error affects the findings is that it influences the tests of significance. Even if the evaluator can be confident that the reliability/validity problems are the same for both groups, the fact remains that reduced reliability or validity

will result in an underestimate of the significance level for the true differences between the groups. The value of f or t or Z or a correlation coefficient cannot reach its maximum unless the measurement is perfectly reliable. Thus, unreliable data and/or data with low validity result in an underestimate of the magnitude of these statistics. The maximum correlation coefficient that can be achieved is estimated as the square root of the product of the reliabilities:[6] $r_{max_{ab}} = \sqrt{rel_a \cdot rel_b}$. Although similar formulae are not available for tests of significance, the effect can be demonstrated by using the value of Z for tests of differences in proportion.

Consider the data in Table 2, where the true scores are in the upper portion and the observed measures (re-contacts) are in the lower part of the table. With a sample of 100 in each group and perfectly reliable data, the Z score for the data in the upper portion of Table 2 would be 3.08 (significant at .002). If 50 percent of the youths who actually commit subsequent offenses are not returned to the court, the value of Z drops to 1.98 (significant at .05). If 75 percent are not caught, the value drops to 1.34 (significant at .18). But if 90 percent of the youths who actually commit subsequent offenses are not returned to the court, the value of Z drops to .685 which has a significance level of .49.

General Principles in Approaching a Measurement Problem

The following are some general principles that evaluation researchers might find useful in approaching any type of measurement problem:

1. Identify the broad concept that the project is trying to have

TABLE 2

HYPOTHETICAL DATA ON TRUE SCORES & ACTUAL OBSERVATIONS[1]

Measurement	Percent with Subsequent Delinquent Offenses	Percent with NO Subsequent Delinquent Offenses
TRUE SCORES		
Experimental Group	20%	80%
Control Group	40%	60%
COURT CONTACT MEASURES		
Experimental Group	10%	90%
Control Group	20%	80%

[1]Court contact figures are based on an assumption that half the youths who commit subsequent delinquent offenses are returned to the court and half either are not caught or, if caught, are not referred to the court.

an effect on or reasonably could be expected to have an effect on and define it in its best, most accurate terms.

2. Get the best possible operational measure of the concept. The evaluator should attempt to measure the concept in the best way possible, but with event/behavior data it is often impossible to have a perfect fit between the variable and the concept. With recidivism data, however, it generally would be the case that measuring closer to the event itself would produce more valid and reliable results than measuring after conviction, for example. This would be true so long as it remains the case that there are more people committing offenses without being caught than there are people being caught who did not commit offenses.

3. Use multiple indicators of the concept whenever possible.

4. Assess the types of measurement problems and the factors that would produce differences between the true (but measured) scores and the actual scores.

5. Determine whether the problems affect both the control and treatment groups in the same manner and, if not, which one is favored by the reliability or validity problems. Results from the study should be interpreted with these types of problems in mind.

6. If the measurement problems affect both groups in the same way, be aware of the fact that tests of significance are conservative estimates (underestimates) when the data are less reliable and/or less valid.

FOOTNOTES

1. See Hubert M. Blalock, "The Measurement Problem: A Gap Between the Languages of Theory and Research, " in Hubert M. Blalock, Jr. and Ann B. Blalock (eds.), METHODOLOGY IN SOCIAL RESEARCH (McGraw-Hill Book Co., 1968), pp. 5-27.

2. Robert Martinson, "What Works?--Questions and Answers about Prison Reform," THE PUBLIC INTEREST, 35, 22-54.

3. For an excellent discussion of reliability and validity in psychological measurement, see the papers contained in William A. Mehrens and Robert L. Ebel, PRINCIPLES OF EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT: A BOOK OF SELECTED READINGS (Rand McNally & Co., 1967).

4. G. Frederic Kuder and Marion W. Richardson, "The Theory of the Estimation of Test REliability, in William A. Mehrens and Robert L. Ebel, PRINCIPLES OF EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT: A BOOK OF SELECTED READINGS (Rand McNally & Co., 1967).

5. Adam Przeworski and Henry Teune. THE LOGIC OF COMPARATIVE SOCIAL INQUIRY (Wiley-Interscience, 1970), p. 103.

6. David Magnussion. TEST THEORY (Addison-Wesley, 1967). Also see William A. Mehrens and Robert L. Ebel (eds.), PRINCIPLES OF EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT (Rand McNally & CO., 1967).

SECTION 2I

MEASURING CHANGES IN THE CRIME RATE*

Abstract

Three kinds of data can be used to measure changes in the crime rate: official (reported) incidences of crime, two or more victimization surveys conducted at different points in time, or one victimization survey covering several months in the recall period. This paper discusses the problems with each approach. In general, the official statistics would be better unless there are reasons to believe that the policy being evaluated altered the reporting of crimes by crime victims to police. Two victimization surveys produce only a pre-post design and, in addition, are difficult to compare if different interviewing procedures were used. One victimization survey cannot be used to measure crime trends because of the problems of forgetting and telescoping.

---

*This paper is excerpted from Anne L. Schneider, "Measuring Change in the Crime Rate," Oregon Research Institute, 1975.

MEASURING CHANGE IN THE CRIME RATE

The Problem

Policy analysts and evaluators of the criminal justice system are confronted with a major dilemma if they wish to examine the performance of the system as a whole in relation to crime reduction. The most widely available performance measures for criminal justice systems are the Uniform Crime Reports published since the 1930s by the FBI. Although these data exist for many areas and many time points, they almost certainly are not reliable indicators of the magnitude in "total" crime. (The term "total" crime refers to both the reported and unreported offenses of a particular type.) The best available alternative data are survey-generated estimates of victimizations, but these data are available for so few time points and so few areas that the more acceptable types of quasi-experimental designs for use in policy analysis or evaluation cannot be used.

The purpose of this paper is to describe problems in measuring trends in the crime rate with official crime data and with victimization survey data.

Use of Official Data to Measure Trends in Crime Rates

Two major problems threaten the accuracy of the official police estimates of changes in the crime rate.[1] One problem is that the procedures used by the police departments to "produce" the official statistics for the Uniform Crime Reports are subject to change over time and the resulting estimates of crime reflect such changes in policy. Crime

waves have been made to appear and disappear through policy decisions concerning how incidents reported to the police are counted and classified.[2] The impact of policy changes on official crime estimates has been studied rather extensively and the results have been dramatic enough that some researchers have concluded that the official statistics are worthless in the evaluation of social policies.[3] Seidman and Couzens, for example, document that the substantial decline in crime for Washington, DC, in 1972 was almost certainly a result of changes in the method of classifying and counting incidents. Spectacular increases and decreases in crime have been observed in other cities as a function of changes in police department personnel rather than as a function of change in the actual crime rate.

The second problem is that the official data upon which the UCR are based contain only the incidents known to the police and do not include incidents that victims fail to report. Thus, the official data are not a count of the "total" (reported and unreported) crime.

If the percentage of victims who report crimes to the police increases over time the UCR rate will increase accordingly even though total crime may remain the same. Variability in victim reporting would make the UCR unreliable indicators of crime change even if the official data were perfect in every other respect.

Complicating the problem for policy analysts and evaluators is the possibility that crime reduction projects and programs may alter the inclination of victims to report crimes to the police. Thus, the problem of victim reporting is especially serious if one is attempting to determine whether changes in the criminal justice system have reduced

area-wide crime. Improvements in the system may increase victim reporting, which in turn will result in an increased number of incidents coming to the attention of the police. If the researcher uses the UCR as an indicator of total crime change, s/he may erroneously conclude that the program under study was ineffective or even detrimental.

General Impact of Reporting Variability

The rate of change in victim reporting to the police will produce the same rate of change in the official estimates of crime, even if the total crime has remained unchanged. The hypothetical data in Table 1 illustrate the point. In the example, total crime (reported and unreported) is 200 per 1,000 persons both at time one and time two. If the proportion of victims who report incidents to the police is 40 percent at time one and increases to 50 percent at time two, the rate of increase in reporting will be 25 percent (50%-40%=10%; 10%/40%=25). The official data also will show a 25 percent rate of increase. As indicated in the table, the lower the percentage of incidents reported during the first time point, the greater the impact a change will have on the rate of change in the official crime data.

If it is reasonable to assume that the percentage of victims who report crimes to the police has remained stable from one time period to another, then trends in official crime could be accurate indications of trend in total crime. However, if reporting varies, a change in official crime can be attributed either to a change in reporting or to a change in total crime. The critical question, then, is the accuracy of the assumption that the tendency of victims to report crimes will

TABLE 1

EFFECTS OF CHANGE IN REPORTING ON OFFICIAL CRIME RATES

IF TOTAL CRIME REMAINS THE SAME

(a) percent reported to police t_1	(b) percent reported to police t_2	(c) rate of change in reporting $\frac{t_2 - t_1}{t_1}$	(d) total crime rate t_1 & t_2 (per 1,000)	(e) official crime rate t_1 (ad)	(f) official crime rate t_2 (bd)	rate of change in official crime rate $\frac{t_2 - t_1}{t_1}$
40	50	25%	200	80	100	25%
40	60	50%	200	80	120	50%
50	60	20%	200	100	120	20%
50	70	40%	200	100	140	40%
60	70	17%	200	120	140	17%
60	80	33%	200	120	160	33%
70	80	14%	200	140	160	14%
70	90	29%	200	140	180	29%

remain stable from one time period to another.

Use of Two Victimization Surveys to Measure Change in Crime Rates

One alternative to the use of official crime data would be to conduct two or more victimization surveys and attempt to estimate change in the victimization rate by comparing the results from the two. There are several problems an evaluator will encounter when attempting to compare these victimization surveys, even if they were conducted within the same city. In particular, the results will be comparable only if all of the interviewing procedures, instructions, and so on are exactly equivalent. A mailed survey one year cannot be compared with a telephone survey the second year. The findings from a high quality, well trained and supervised group of interviewers conducting the survey at one point in time should not be compared with a haphazard administration at a different point in time.[3] Furthermore, given the cost of victimization surveying, the evaluator should be cognizant of the fact that even if two surveys are conducted, s/he is left with one of the weakest of all possible evaluation designs: a pre-post, no control group design. Thus, an evaluator should be very cautious before recommending that an evaluation be based on a pre-post set of victimization surveys. Surveys are suitable for measuring change in the crime rate if (1) they are conducted on a regular basis for several years, (2) the technical requirements for survey research are met, and (3) pre-post surveys are done on treatment and comparison areas or treatment and control households.

Use of One Victimization Survey to Measure Change in Crime Rates

There are four sources of bias in victimization data from a single survey that prevent researchers from analyzing the trend during the recall period covered by the survey:

1. Respondents telescope events into the time period which actually occurred prior to the first month which was covered in the survey recall period.

2. Respondents telescope incidents both forward and backward within the recall period, but the net effect is a forward telescoping of events.

3. Respondents forget some of the incidents which occurred and the memory loss is greatest for the most distant months covered in the recall period.

4. The actual month of occurrence cannot always be recalled by the respondent and the tendency to forget the date is most apt to occur for incidents during the most distant months.[4]

The combined impact of these biases is such that one should always expect victimization survey data, when analyzed on a month-by-month basis, to show that the victimization rate increased during the time period covered by the recall period.

It should be noted, however, that these biases tend to be uncorrelated with characteristics of individuals and it is likely, therefore, that bias in the data would be the same for a comparison and a treatment area of a city or for control and experimental households.[5] Thus, given a proper type of design, the evaluator could utilize trend data from a single survey.

FOOTNOTES

1. Two other major problems with the UCR are not discussed in this article, but should be noted. One concerns victimless crimes--sometimes called "satisfied customer" crime--such as narcotics peddling, prostitution, gambling, and so on. These are not included in the UCR index. A second category of poorly tabulated crime pertains to employee pilfering, shoplifting, and some types of fraud. These are rarely reported as they occur and are generally detected during periodic accounting procedures. Thus, the UCR give no indication of level or change in these. A second problem concerns the fact that the method of classifying crimes for UCR reporting is not directed at capturing what is needed for a measure of societal deviance. Also see S. Wheeler, "Criminal statistics: A reformulation of the Problem," in JOURNAL OF CRIMINAL LAW, CRIMINOLOGY AND POLICE SCIENCE, 1967, 58, 317-324.

2. A. Biderman and A.J. Reiss Jr., "On Exploring the 'dark figure' of crime," in ANNALS OF THE AMERICAN ACADEMY OF POLITICAL AND SOCIAL SCIENCE, 1967, 374, 1-15. Also see D. Seidman and M. Couzens, "Getting the Crime Rate Down: Political Pressure and Crime Reporting," in LAW AND SOCIETY REVIEW, 1974, 8, 457-493.

3. It is reasonable to believe that poor interviewers or interview procedures will discourage the recall of crimes by respondents and that the type of events most likely to be forgotten or assumed to be

irrelevant for the interviewer to bother with are the trivial ones. Thus, the victimization rate would be too low and, since trivial incidents are less likely to be reported to the police, poor interviewing procedures will over-estimate the percentage that are reported to authorities.

4. See Anne L. Schneider, William R. Griffith, David Sumi, and Janie M. Burcart, "The Portland Forward Records Check of Crime Victims: Final Report," Institute of Policy Analysis, December 1977.

5. Ibid.

SECTION 2J

MEASUREMENT STRATEGIES FOR DETERMINING CITIZEN POLICY PREFERENCES*

Abstract

Several issues and measurement strategies for assessing citizen policy preferences are discussed in this paper.

*This paper is a revision of materials prepared by L.A. Wilson II and Anne L. Schneider in response to a technical assistance request during the Model Evaluation Program.

MEASUREMENT STRATEGIES FOR DETERMINING CITIZEN POLICY PREFERENCES

Introduction

Victimization surveys of the general public normally have been conducted for the purpose of measuring the "true" (reported and unreported) crime rates. Survey research, however, is useful to criminal justice planners, analysts, and evaluators in several other ways. The purpose of this paper is to discuss some of the issues and approaches for measuring citizen preferences concerning the priority that should be given to public safety vis a vis other types of public services. These same issues and techniques could be used to measure citizen preferences about the priority that should be given to different types of crimes and/or different types of criminal justice system activities.

Choice of Referent

Questions about policy preferences can focus on the functional expenditure categories found in most city budgets, such as police, courts, parks, schools, and city transportation. The survey could measure the relative importance of these, the amount of funds that "should" be spent on each, satisfaction with current service levels provided, and so on. One advantage of this approach is that the results can be compared rather quickly and easily with the actual distribution of city resources, since the referent in the question corresponds to an existing budget category. A problem with using functional categories as the referent is that citizen perceptions of these probably are not as sharp nor as personal as

they would be for some of the other choices.

A second choice of referent is to concentrate on goals and values that correspond (more or less) to the range of services provided by the public agencies. This would include such things as safety from different types of crimes, safety from fire, rapid response from police or firemen in times of emergency, education, environmental quality, convenient transportation, access to shopping, recreational facilities, etc.

A third choice is to measure citizen preferences concerning the specific means of achieving different goals. For example, citizens might be asked to judge the comparative importance of patrol cars, community crime prevention, and speedy trials in achieving the general goal of reduced crime.

A common problem in measuring citizen preferences is that the means/end distinction is ignored. Citizen responses to questions about their preferences concerning the proportion of resources that should be allocated to the police, schools, or parks can reflect either the relative importance of the various goals or the respondent's perception of the likelihood that the agency could achieve that goal. For example, a respondent might believe that crime reduction is the most important goal, but indicate in response to a question that the bulk of the resources should go to education and recreational programs. Underlying the response, perhaps, is the notion that education and recreation are more effective in crime prevention than are the police and courts.

Probably the most relevant point is that the questions should not confuse means and ends. If one is interested in information about the relative importance of various social goals, then the questions should

focus on goals, not on agencies or agency activities. If the interest is in citizen perceptions about the effectiveness of different methods for achieving the goals, then this should be made clear in the question and the method as well as the goal should be mentioned.

Dimensions of Preferences and Perceptions

Most of the previous research that has been done on citizen preferences and policy focuses on satisfaction with the current level of service (or operating procedures), the priorities that should be given to different policy areas, social problems, or social goals, and (rarely) the value of incremental changes in amount of the service or goal. The choice among these depends mainly on the purposes of the study. If one is interested in comparing public preferences concerning priorities or distribution of resources with the actual (official) priorities or distribution of resources, then "satisfaction" is not as good a choice as the others. Even though a person might be more satisfied with one type of service than with another, this does not provide information on which should be given higher priority.

Questionnaire Construction and Measurement Strategy

Likert Scales: One of the most frequently used methods of measuring citizen preferences with surveys is to use Likert-type items in which one goal or problem is asked about in each question. For example, satisfaction with existing services often has been measured by asking whether the respondent is "very satisfied, somewhat satisfied, somewhat dissatisfied, or very dissatisfied" with educational services, police services,

streets, etc. Clearly, this method leaves much to be desired if one wants to compare the service levels or if one wants to determine the citizen's preferences concerning the priorities that the government should give to different policies.

Ranking: Ranking of problems (in terms of their severity) or of goals (in relation to their value) also is used at times to measure citizen priorities. The problem with this procedure is that one cannot compare citizen preferences with government actions unless government officials also are interviewed and asked to respond to the same questions.

Paired Comparisons: Paired comparisons would be a useful approach for determining the priority citizens give to different problems or goals, particularly if one used unfolding analysis to generate an interval-type scale.[1] The problem with paired comparison, in field research, is that the number of questions generated can be extremely high since all possible pairs need to be included in the questions. Pre-tests of an instrument that asked for the comparative importance of police protection, fire protection, education, and two other service areas generated ten pairs and it took twenty to thirty minutes for respondents to answer the questions. Part of the problem was that there were ten separate questions, but in addition the interviewers reported that respondents simply could not determine whether "police protection or fire protection" was more important to them. The zero-sum context of the responses probably contributed to the length of time required to administer that portion of the survey.

"Budget Pie": Another possibility that would be worth considering (for personal interviews) is a "budget pie" type of question. The basic

idea is to provide the respondent with a list of problems (or goals, agencies or activities) and have them "slice" a budget pie to show the proportion of total resources that they think should be distributed to each problem (or goals, etc.). The circle should have a center point with a line at the 12 o'clock position to provide the respondent with a starting point and probably should be relatively large to insure that the person has room to include everything.

Obviously this type of question would be impossible to use in a telephone interview. Another option would be to have the respondent simply list the proportion of resources that should go to each of several policy areas. The problem with this is that the person will have a hard time getting the total to add up to 100 percent, whereas in a confined space (such as a budget pie) it is easier to conceptualize how much is left to distribute after each response has been made.

The budget pie approach specifically asks for preferences concerning the allocation of existing resources and, therefore, is easily compared to actual governmental allocations. One of the problems with asking persons to indicate the comparative "importance" of various problems or goals is that "importance" is not very specific nor very clear. Asking for preferences concerning allocation of resources is not as ambiguous. The budget pie type of question also can be used to determine how persons would like to allocate additional resources if some should become available.

Magnitude Estimation: Another measurement strategy that might be appropriate, especially for telephone surveys, is magnitude estimation.[2] The basic idea is to give the respondent one referent, assign it a "score,"

and then ask for the "score" that the respondent would give to each of several other referents. (The severity of different types of crimes often is assessed in this way.) It might be possible to use this strategy to determine the comparative dollar amounts that the respondent thinks should be given to solve certain types of problems.

Supply and Demand Curves: A problem with all of these procedures is that none of them provides information concerning how much the individual would be willing to pay to achieve various proportions of certain social goals. The budget pie procedure results in estimates of citizen preferences concerning allocation of resources and these preferences can be compared to actual allocations or could be used to provide guidance concerning future resource distribution. The magnitude estimation procedure might be useful for determining the ratio that citizens would prefer concerning allocation of resources across different types of policy areas. But the economists would prefer a measure that indicates the public demand for services for each of several different prices because this would permit an estimate of the optimal amount of resources that should be given to each service--not just the proportionate share.

One method that might be used to ask these kinds of questions would be to ask the respondent how much they would be willing to pay, in taxes (presuming that everyone else paid their fair share), in order to achieve some portion of a goal. One of the problems with this is simply in figuring out how to phrase the questions. Information is not available as to the quality of responses that would be obtained if someone simply asked the respondent, "How much more would you be willing to pay in taxes (presuming that everyone else paid their share as well) in order

to reduce the number of burglaries in the city from 20,000 to 15,000? How much more would you be willing to contribute if you were certain that the result would reduce the number of burglaries to 14,000?"

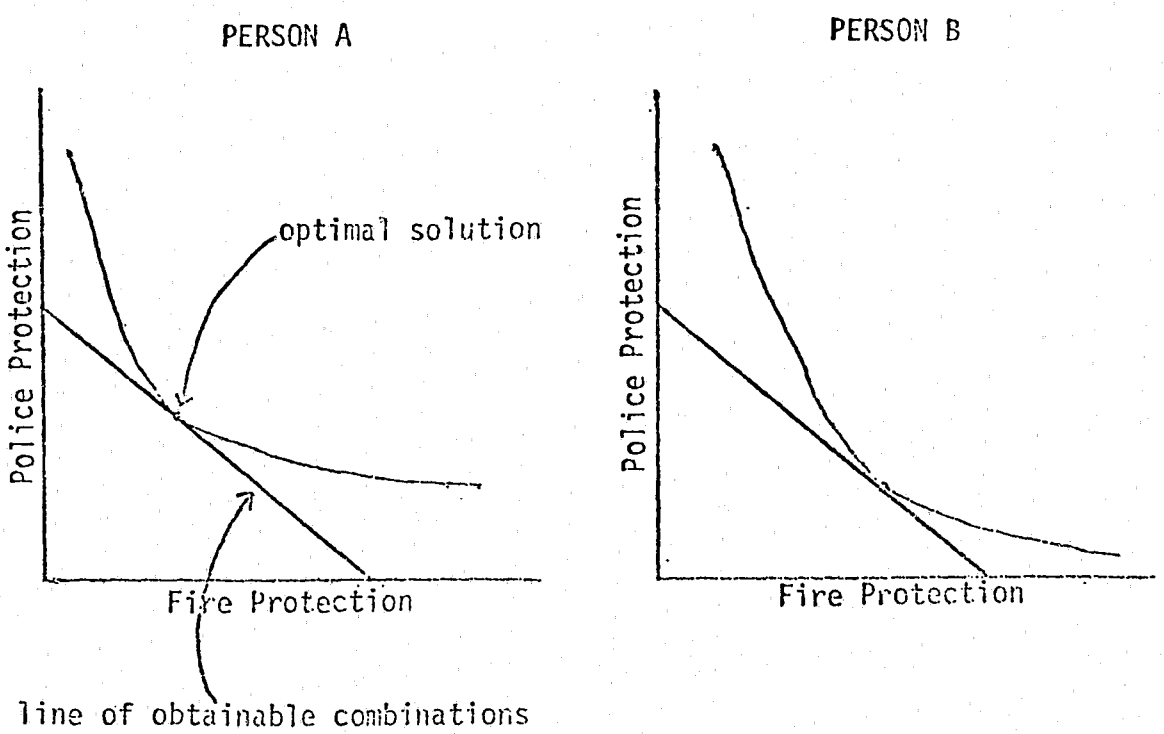
An interesting twist on this general approach would be to tell the respondents that they have been given a specified amount of a particular desired goal and then ask them how much they would sell it for. This procedure would work well for measuring the value of certain types of things. For example, one could tell the respondent that they have been given a free parking permit for the downtown area and then ask them how much someone would have to offer before they would sell it. The value of police patrols might be assessed in this way, but it would be difficult to think of a plausible way of asking these questions (for example) in relation to rape prevention.

Indifference Curves: Another approach would be to use survey data for the purpose of constructing indifference curves and from these estimate the tradeoffs that citizens would prefer between different services.

Since a variety of public services are offered, it makes no sense to ask an individual how much of any one service in isolation from others s/he wants. Conversely, it is impossible to ask the respondent to deal with all public services at once, deciding how much would go to each type of service in competition with all other services. Even if possible, it should be assumed that such responses would be quite errorful. The method of approaching this problem would be one which permitted the definition of an indifference curve representing the tradeoff between different types of services which would be optimal, in the eyes of the respondent.

The use of indifference curves in looking at individual preferences

for the provision of different types of public services offers a number of advantages over other methods. First, all expressions of preference are presented in terms of comparisons of the utility of at least two different services. Second, there are constraints in the number of goods and services which can be provided. That is, "real world" constraints in the amount of all goods and services which can be provided are explicitly stated and dealt with. Third, it is possible to identify optimal solutions for individuals and groups and assess the extent to which these solutions vary from existing conditions. To illustrate some of the points made above, consider the following indifference curves:



The line of obtainable combinations places the real world constraint upon the decision which is to be made. That is, a known ratio of one public good to another can be provided for the same amount of money. This ratio runs, in this example, from all fire protection/no police protection (far right of each diagram) to all police protection/no fire protection (far left of each diagram). Between the right and left extremities an infinite range of ratios of one good to another are portrayed.

When an individual is asked the simple question, "How much police protection do you want?" (or some variety of this with no comparison or constraint imposed), one is permitting the line of obtainable combinations to move upward and to the right. That is, more of all goods are being provided at any point above zero when the line of obtainable combinations moves in this direction.

The optimal solution is reached when the indifference curve intersects with the line of obtainable combinations. A suboptimal solution exists when the actual amount of the goods (or combination of the goods) produced is either above or below (on the line of obtainable combinations) this point of intersection.

Although of obvious interest to the researcher, the description of the individual's indifference curve is not as important as the knowledge of whether the individual perceives the amount of the good currently produced as the amount of the good desired when compared with the production of another good. That is, has an optimal solution been reached in the production of two public goods or services? Additionally, we should seek to discover the extent to which the combination of goods produced is suboptimal.

The measurement problem can be dealt with through the use of paired comparisons. The items to be used might be of the following nature:

1. Assume for the moment that you are presently receiving 100 units of police protection and 100 units of fire protection. If you were able to exchange some of the police protection for some of the fire protection (or vice versa), which one would you decrease and which would you increase? (Probe, if necessary.) Would you give up a little fire protection to have a little more police protection, or would you give up a little police protection to have a little more fire protection?

   increase_____(A)_____ decrease_____(B)_____

   How much of the 100 units of __(B)__ would you give to __(A)__?

2. Assume for the moment that you are presently receiving 100 units of police protection and 100 units of parks and recreation. If you were able to exchange some of the police protection for some of the parks and recreation (or vice versa), which would you increase and which would you decrease?

   increase_____(A)_____ decrease_____(B)_____

   (Probe, if necessary, in same way as above, and then ask:)

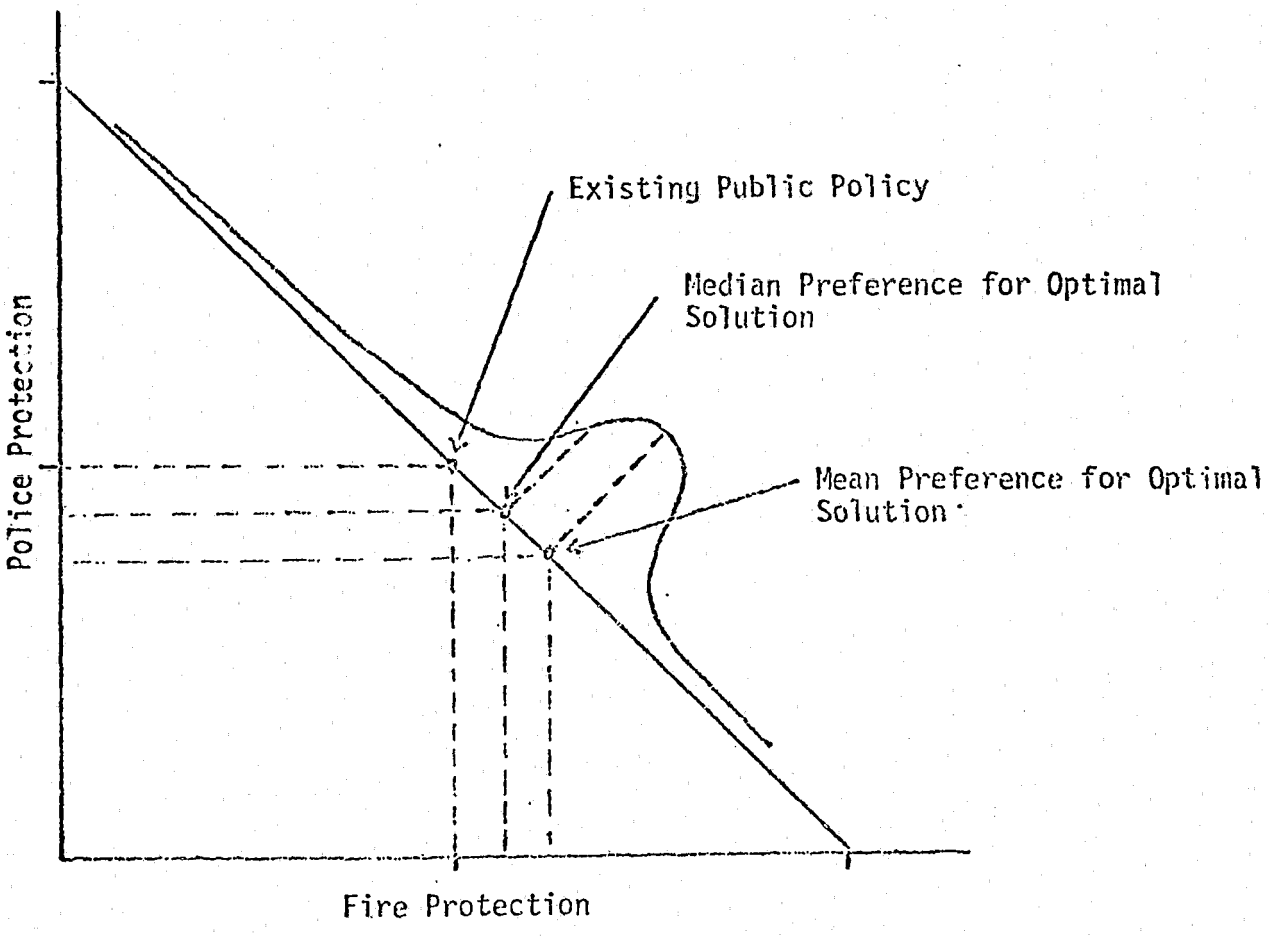
   How much of the 100 units of __(B)__ would you give to __(A)__?

Items such as those listed above could be repeated for as many public services as deemed of interest. The use of such questions would yield a collection of points in Cartesian coordinates. Each of the points represents the particular combination of public goods and services that reflects the intersection of each individual's indifference curve with the line of obtainable combinations--each person's view of an optimal solution.

By setting up the problem in this way, we are making the assumption that the current public policy is represented by 100/100 in any combination of goods or services. This may or may not represent the most optimal solution when one aggregates citizen preferences. The extent to which the measure of central tendency for the survey sample

diverges from the 100/100 solution would represent the extent to which public policy is out of synch with citizen preferences.

A possible outcome of this type of analysis is portrayed below:



This method of analysis would provide both individual level data, against which could be measured conformance of public policy with aggregate preference. A number of assumptions are made in making use of this strategy. Some of the more important ones are listed below:

1. In using this approach we are focusing upon outputs of public policy. That is, we are looking at units of output, not units of input.

The problem with looking at units of input, such as dollars, is that the individual is then asked to translate those dollars into units of output. Since this relationship is not known, we should explicitly focus upon units of output.

2. In the same vein, by assuming that the individual is receiving 100 units of each good at the present time, we are standardizing the amount of the good being received. It might be helpful to view the 100 figure as a percentage of the good that is being received. Clearly, the individual is receiving 100% at this time (the 100% figure will obviously differ across individuals in terms of the actual amount of the good they are currently receiving).

3. The approach also assumes that we are interested in the extent to which the current provision of public goods and services is reflective of the optimal outcome as perceived by the aggregate of citizens.

4. Finally, it assumes that we can, with some precision, specify the relevant and inclusive public goods and services that are to be compared.

FOOTNOTES

1. See Robyn Dawes, FUNDAMENTALS OF ATTITUDE MEASUREMENT (John Wiley & Sons, 1972).

2. See Robert L. Hamblin, "Social Attitudes: Magnitude Measurement and Theory," in Hubert M. Blalock (ed.), SOCIAL STATISTICS (McGraw-Hill Book Com., 1972).

SECTION 3

TECHNIQUES FOR IMPROVING THE UTILITY OF EVALUATION FINDINGS IN PLANNING, PROJECT OPERATION, AND DECISION MAKING

Overview

The five papers in this section are designed primarily for planners, project directors, and other decision makers. Nevertheless, they should be studied carefully by evaluators in order to achieve a common understanding (and a common set of terms) about evaluation.

The general thrust of the papers is to pinpoint the role of evaluation in the planning and decision making processes, to aid the consumers of evaluation in determining the appropriate questions to be asked in an evaluation, and to specify how one can determine the appropriate criteria for the "success" of a demonstration project.

SECTION 3A

AN INTRODUCTION TO EVALUATION FOR PLANNERS AND DECISION MAKERS*

*This paper was written by Peter R. Schneider and Anne L. Schneider.

AN INTRODUCTION TO EVALUATION

FOR PLANNERS AND DECISION MAKERS

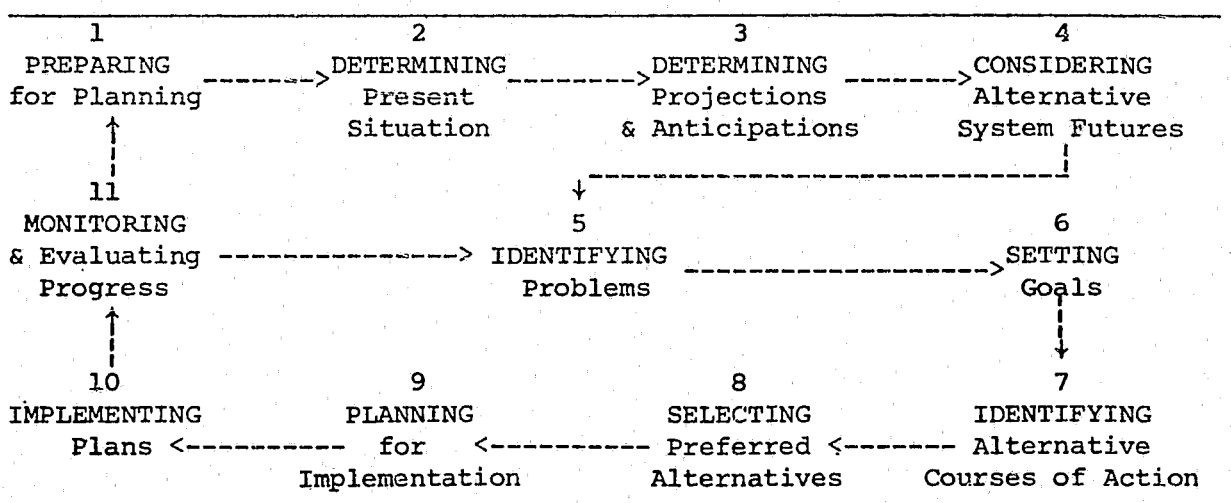
Introduction

The purposes of this paper are to pinpoint the role of evaluation in the criminal justice planning process, to acquaint planners and other decision makers with the different types of evaluations and with the kinds of information evaluation provides, and to discuss the roles of evaluators, planners, and other decision makers in selecting the questions to be answered and the methodology to be used.

The General Planning Process Model diagrammed in Figure 1 shall serve as our point of departure.[1]

FIGURE 1

GENERAL PLANNING PROCESS MODEL



In this model, evaluation (defined both as the monitoring of a project's activity and the assessment of its impact) is perceived as being

central to the planning process. First, evaluative information can assist planners in determining what is currently happening within the criminal justice system, what is likely to happen during the planning cycle, and what changes, if any, can be expected (steps 1-4). Second, evaluators can assist in the precise identification of problems, the setting of measurable goals, and the determination -- again based on evaluative information -- of what types of strategies may be transferable from other settings (steps 5-7). Third, evaluators can contribute to the policy-making process by appraising specific projects and helping to order them on a list of priorities for implementation (steps 8-9). Finally, evaluation of the projects which have been implemented provides project directors, planners, and decision-makers with the feedback which is necessary for continuation and/or modification decisions.

Types of Evaluations

A number of different types of evaluations can be identified. Perhaps the best-known typology in the literature on evaluation is the distinction made by Michael Scriven between formative evaluations and summative evaluations.[2] Formative evaluations are oriented toward the "front-end" of the planning process and are generally conducted for the purpose of helping to develop new programs or contributing to decisions about program installation. The activities undertaken in a formative evaluation roughly correspond to those involved in steps five through nine of the General Planning Process Model, and would include such things as determining the need or demand for a program, assessing the likelihood of its effectiveness, and appraising the adequacy of resources for carrying it out. Summative

evaluations, on the other hand, are oriented toward the "back-end" of the planning process and are conducted for the purpose of assessing the overall effectiveness of a given program or project and contributing to decisions about its continuation, expansion, or transferability to a different environment.

The Law Enforcement Assistance Administration, in an effort to standardize terminology for criminal justice evaluations, issues and occasionally updates guidelines defining the different types of evaluations. In the current version of the guidelines, a distinction is drawn between monitoring and evaluation. While both compare the results or achievements of a project with its intended objectives, evaluation involves a more intensive analysis than monitoring and attempts to verify that the achievements or results of a project are, in fact, attributable to the project's activities. For example, the operation of a project designed to reduce juvenile recidivism within a pre-selected target population may indeed coincide with a reduction in new offenses among the youth participating in the project, and a monitoring report would be required to do little more than determine whether the amount of the reduction met the project's objectives. However, an evaluation would be required to demonstrate whether the activities of the project were responsible for the lower rate of new police contacts, and that alternative explanations, such as increasing maturity and/or greater skill in escaping detection, could be eliminated.

There are at least two problems with the existing LEAA guidelines defining the types of evaluations: First, there is ambiguity as to what constitutes the "results" of a project. In an anti-burglary project, for example, one "result" might be that the police sent pamphlets to 20,000

households; another "result" would be that they engraved property in 5,000 homes; still another result might be that burglary in the targeted sections of the city actually declined, and a fourth result might be that burglaries increased in other areas of the city due to the displacement effect. There is no distinction, in other words, between activity levels, results and broader range outcomes. A second problem with the current definitions is that the major difference between monitoring and evaluation concerns the intensity of the effort and the kinds of research strategies employed, and not the purposes for which the evaluation is undertaken. Only one purpose, in fact, is recognized in the guidelines: to compare project achievements with project objectives.

A training course currently being developed under LEAA auspices contains another typology of evaluations which appears to correct the major deficiencies of the LEAA definitions.[3] In this scheme, a criminal justice project is conceived of as a system consisting of inputs (resources, guidelines and operating procedures); activities (those things the project and its personnel do); results (the initial consequences of the activities); and outcomes (the long-range, socially relevant consequences of the project.)[4] The system should contain a feedback loop through which the results and outcomes of a project impact upon the operation of the project and act as additional inputs.

The LEAA training course, therefore, differentiates among the types of evaluations according to the point in the system where the final performance measure is taken. It the evaluation attempts to link one or more outcomes in a causal fashion back to results, activities and inputs, it is referred to as an <u>impact assessment</u>. If the evaluation focuses on results

rather than outcomes, and the results are linked to activities of the project, it is called a process evaluation. Finally, an evaluation in which activities are linked in a causal fashion to inputs is called monitoring.

Although this typology is quite adequate for some purposes, it does not indicate the kind of comparisons that are going to be made and, therefore, is inadequate for the purposes of actually formulating the questions that the evaluation would seek to answer. Thus, a second dimension has been developed that will further expand our understanding of the various types of evaluations which could be conducted.

A "black box" evaluation is one in which the entire project is compared with some alternative strategy of achieving the same objectives.

A "project component" evaluation is one in which a specific component of the project is compared with some other component within the same project.

A "multiple linkage" evaluation is one in which the activities are related to the outcomes through one or more intervening linkages. For example, one could propose that a crisis intervention program for status offenders would reduce the proportion held in detention and that this, in turn, would reduce recidivism. Alternatively, one could propose that a crisis intervention program would reduce recidivism because it is a more effective type of counselling and that a reduction in detention (even if it occurred) would be unrelated to recidivism. A multiple linkage evaluation is one that would test which of these hypotheses is correct.

What Type of Evaluation is Needed?

The type of evaluation that is needed depends on the questions that one wishes to answer with the evaluation. In turn, the questions that need to

be answered depend on the "developmental phase" of the project and on the kinds of decisions made by potential users of the evaluation findings.

Some would suggest that all projects should have each type of evaluation (monitoring, process evaluation, and impact assessment) all the time, but the expense of evaluation is such that choices must be made. In order to make these choices, it is useful to view a project as going through four stages, beginning with its initial funding and implementation in a community and continuing through its achievement of a maximum level of efficiency.

At Phase I, when the project has just been funded, the critical questions are whether the resources and guidelines (inputs) are producing the desired level of activities and whether the internal operating procedures of the project are contributing to the achievement of these activity levels. Ideally, the cost effectiveness of each project component would be ascertained. Cost effectiveness refers to more than simply whether the project is achieving its specified activity levels, but whether the per unit cost of each project activity is reasonable, when compared against other methods of producing the same activities. Monitoring of a project should provide answers to these questions.

At Phase II, the project has been implemented, its activities are underway, and some client or areas or other parts of the criminal justice system are receiving the services or "treatments" of the project. At this point, the critical questions are whether the activities are producing the desired results (initial consequences or short-term effects), and whether any unexpected or undesired results are occurring. Again, the best

procedure would be to examine the cost effectiveness of each project activity in achieving these results by comparing different project activities with each other or by comparing the whole project with some alternative method of achieving the results. Process evaluation is most appropriate at this phase of project development.

At the third phase, the project has been implemented for a sufficient length of time that it is reasonable for some broader-range social consequences to have occurred. For some types of projects, these should appear almost as soon as the initial results (within a few weeks or months) whereas with others it may take considerably longer (a year or so). The key questions of concern during this phase are whether the strategy or theory underlying the project is sufficient to produce the desired social consequences, such as reducing crime or increasing the quality of justice. Virtually all projects have either an explicit or implicit rationale (theory) which makes it reasonable to believe that the activities will produce the desired results and the desired results, in turn, will produce the expected impact on the problem(s) that the project was supposed to solve or ameliorate. The impact assessment, which determines whether the activities/results are producing the desired outcomes, is appropriate at this phase of the project.

As before, the ideal impact assessment would not only verify that the theory of the project is working properly (e.g., the activities produced the desired results which, in turn, produce the desired outcomes), but also would indicate the optimal level of resources, activities, results, and outcomes by determining whether the strategy used by the project is more cost effective than any other available strategy that could be used by the same or a different project.

The fourth and final phase occurs after the project has become a routinized, on-going part of the criminal justice system. Continuing evaluation (usually process evaluation) is needed to determine whether changes are occurring in the environment or in the project operating procedures that would alter the relationships established in the previous evaluations. For example, if the characteristics of offenders are shifting from professional to juvenile, then previously effective strategies may become ineffective in reducing crime.

The type of evaluation that is needed also depends on the decisions that will need to be based on the findings from the evaluation.

Project directors, initially, will be most interested in examining the various components of the project to determine which resources, rules, or operating procedures are the most cost effective in producing activities; which activities are most cost effective in terms of immediate results; and which of the activities or results are the most cost effective in relation to the outcomes. This information can be used to eliminate some activities or re-allocate resources among the activities.

Agencies that are responsible for allocating funds to many projects, some of which have similar purposes, will be interested in determining which projects in comparison with other projects are most effective in using resources, maintaining cost effective activity levels, producing results, and achieving cost effective outcomes.

If the questions of concern to decision makers include an assessment of project effectiveness in utilizing resources, achieving results, and producing the desired outcomes, then an impact assessment should be conducted but it should have a monitoring and process evaluation component.

It is important to note that the planning required for the more complex evaluations, especially impact assessment, must begin even before the project is implemented. Thus, if an impact assessment is needed, it should be implemented before Phase I of the project development cycle.

Evaluation Methodologies

Evaluation is a type of research that attempts to establish a causal linkage between the project (or some aspect of the project) and one or more consequences of that project or project component. Impact assessment requires that the desired outcomes not only occur, but that the role of the project in producing those outcomes be clearly identified. Process evaluation requires not only that the results occur, but that these can be attributed to the project rather than to some other factor external to the project. Even in the newer LEAA definition of monitoring, the method used should establish the linkage between inputs and activities in order to demonstrate that the activities would not have occurred without those resources and other inputs.

The importance of establishing a causal relationship should not be underestimated. Consider a situation in which an evaluation report claims that a particular project (costing $100,000) reduced burglaries that would have resulted in $200,000 loss to victims. The project would seem to be very cost effective in comparison with the way the system operated without the project (it "saves" $2 for every $1 spent). This conclusion, however, would not be warranted unless the evaluator can demonstrate that this entire reduction in burglaries (and loss from burglaries) was due only to the project and not to other factors. If the

reduction would have occurred anyway, then the project is not cost effective at all. If part of the reduction is due to other parts of the system, then the cost of these should have been included and the cost effectiveness for the entire system should have been assessed.

One of the key elements in establishing a causal relationship involves the evaluation design that is to be used. The most common typology of evaluation designs distinguishes among experimental designs, quasi-experimental designs, and pre-experimental designs.

1. Experimental Designs. In an experimental design some of the clients (cases, areas, and so on) who are eligible for the "treatment" are randomly selected to be in a control group that does not receive the treatment or receives a different type of treatment. This design generally is implemented in field situations by first identifying a group of persons (or areas or cases) that are eligible for the treatment or are in need of the treatment and then choosing some to receive it and others not to. This becomes a "denial of services" only if the treatment clearly is better than the alternative used for the control group and only if the level of resources for the treatment is sufficient to handle all those who are eligible or who need the intervention. In most situations. an experimental design that is properly implemented will insure that the true effectiveness level of the treatment can be ascertained and that consequences of the project can be properly separated from outcomes that would have occurred anyway.

2. Quasi-Experimental Designs. Quasi-experimental designs require no random assignment, but the evaluator must have other groups, areas, or cases that are relatively equivalent to those receiving the intervention

and must compare these with the project groups (or areas or cases). There are several quasi-experimental designs, including interrupted time series and pre-post comparisons of the treatment group with a relatively equivalent group of persons who received some other type of treatment or no treatment at all. These designs will not insure that the apparent outcomes of a project can clearly be separated from outcomes which would have occurred anyway, but some of the quasi-experimental designs are quite good. Part of the difficulty in using a quasi-experimental design is that one may not know until after the data have been analyzed whether all of the alternative explanations for the apparent effects of the project can be ruled out. Thus, the risk of conducting the evaluation and still not knowing the answer to the questions of interest is greater with quasi-experimental designs than with experimental designs. Nevertheless, a skilled evaluator may use several different quasi-experimental approaches and, through the consistency in results, be able to answer most of the relevant questions.

3. Pre-Experimental Designs. The most common type of pre-experimental design is one in which the evaluator compares a "post-project" observation (such as the crime rate) with a "pre-project" observation for the group or area that received the intervention. This design is so weak that virtually no valid conclusions about project effectiveness in relation to results or outcomes can be drawn when it is used.

Although the establishment of causal linkage usually is quite difficult and requires a strong design, there are some situations in which it is not as complex. In monitoring, for example, it is often sufficient to simply document that the resources are being used and the

activities are occurring. There usually are no alternative explanations for why the activities could have occurred. The evaluator will be able to obtain documentary evidence or to observe the project in operation and ascertain that the activities would not have occurred without the project. But when the performance measure is a result or an outcome, there usually are many alternative explanations for why the result or outcome occurred. It is generally quite difficult to obtain valid evidence that the crime rate, for example, declined because of the project and would not have declined without the project.

Interrelationships During Project Planning and Implementation

In order to establish a causal relationship between the project and the consequences of interest, the evaluator must have a strong research design and must have data that are reliable and valid. Although there are other factors involved in producing valid conclusions from evaluation, these two are of considerable relevance to planners, project directors, and decision makers because the evaluator cannot, without the assistance of others, ensure that the data or design will be sufficient to produce scientifically valid answers to the questions of interest.

In order to ensure that the proper data are collected and that the data elements are both reliable and valid, the evaluator should be involved in the development of data collection instruments and procedures even before the project starts. If funds are sufficient, the evaluator should have responsibility for designing instruments, training persons to collect the data, and checking the reliability of the data. In

order to ensure that the design will be sufficient to provide valid answers to the questions of interest, the evaluator should be involved in discussions and negotiations about the operation of the project before it is implemented. If the evaluator is called in months or years after the project starts, it is quite possible that no valid conclusions can be drawn about the causal relationship between the project and its presumed effects because the design is too weak, the data are not reliable, or the relevant data were not collected.

The evaluator also should be involved in the determination of which questions the evaluation will attempt to answer. The extent of evaluator involvement in this determination depends on the situation and on the skills the evaluator has in anticipating the future informational needs of planners, project directors, and other decision makers. If the evaluator is quite skilled at this, then s/he could prepare the initial list of questions that might be important, the estimated costs of the evaluation if it is to produce valid answers to those questions, and the implications of the design used to answer the questions for project operation and data collection.

Summary

In general, it is reasonable to say that evaluation produces information that could be used in planning and decision making, but in order for this to happen, the evaluator should be involved in the development, planning, and implementation of the project itself.

Evaluation can fail to serve the purpose of guiding decisions if the relevant questions are not asked or if the answers provided by the

evaluation are not valid. In some situations, the evaluator's role has been limited to the technical research aspects of producing valid conclusions, whereas the planner's (or other decision maker's) role has been limited to project development or operation. The approach described above involves a series of discussion and negotiation sessions in which the evaluator, planner, project director, and other relevant decision makers clarify the questions to be addressed in the evaluation, the design to be used, how this is to be implemented by project personnel, and the data collection procedures.

FOOTNOTES

1. This model is the one being used by LEAA in its Training Course for Criminal Justice Planners.

2. Michael Scriven, "The Methodology of Evaluation," in PERSPECTIVES OF CURRICULUM EVALUATION (Chicago: Rand McNally, 1967).

3. The training course for criminal justice evaluators.

4. This scheme is only a slight modification of a general systems approach. The major differences are in terminology, the placement of the boundaries around the system (e.g., project), and absence of any explicit feedback loop, and in the lack of emphasis on the environment.

SECTION 3B

A SYSTEMS APPROACH TO EVALUATION*

Abstract

The purpose of this paper is to describe a series of procedures that serve to pinpoint the questions and propositions which should become the focus of an evaluation.

*The information presented in this paper and the terminology used are consistent with the LEAA Evaluation Training Course, but some of the materials here are not used in that course. The author of this paper (Dr. Anne L. Schneider) teaches a module in the LEAA course that focuses on some of the topics covered in this paper.

A SYSTEMS APPROACH TO EVALUATION

Introduction

The fundamental purpose of evaluation is to produce scientifically valid information and conclusions that will be useful to planners, project directors, and other decision makers. Evaluation will meet these objectives only if the questions to be answered by the evaluation are relevant, rather than trivial, and only if the procedures used to draw conclusions are consistent with the standards of social science inquiry.

The purpose of this paper is to describe a systems approach to evaluation that should be useful not only to evaluators, but also to planners and project directors, in their efforts to identify the types of questions that should be addressed in an evaluation.

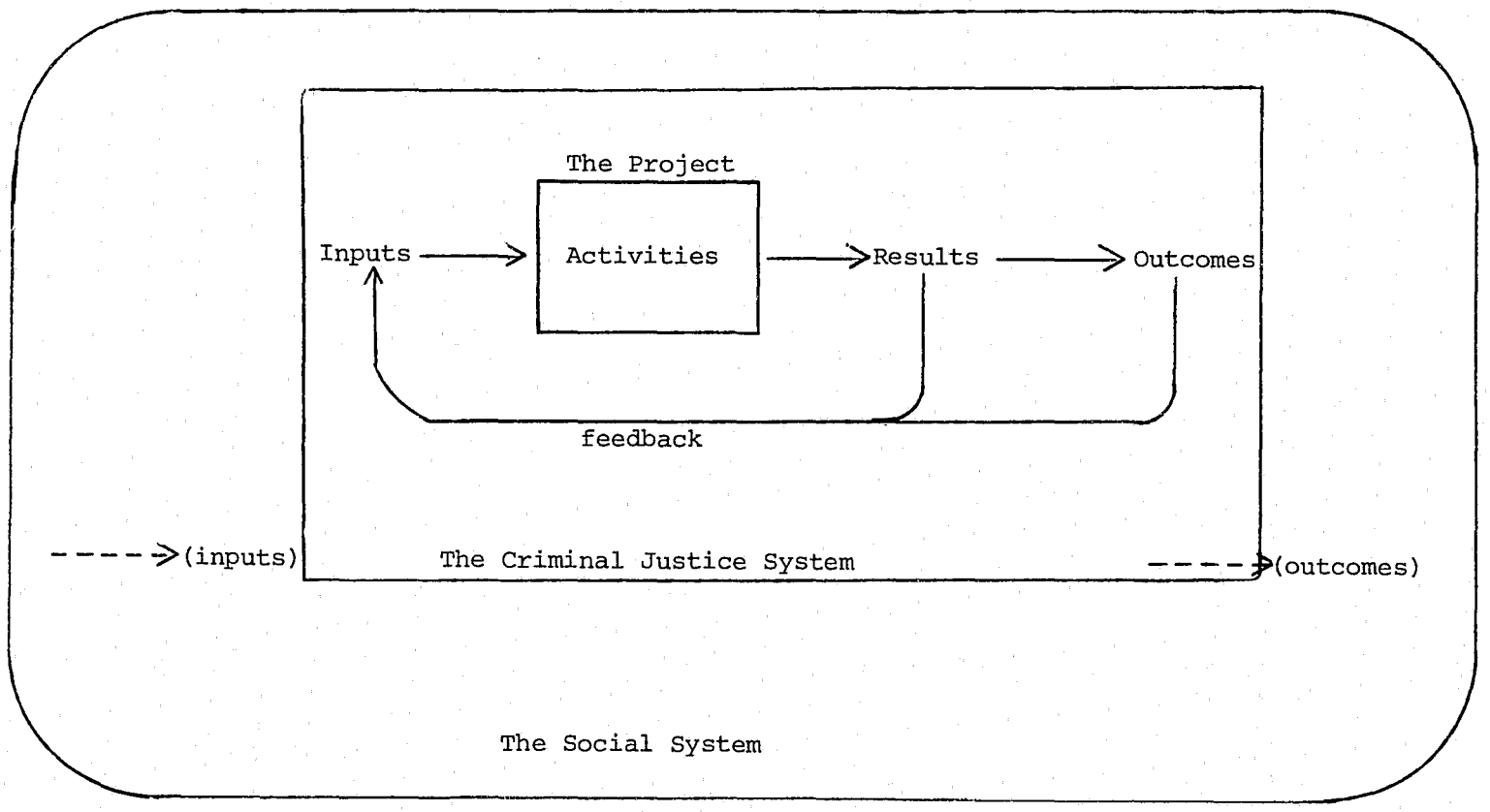
Components of the System

One of the first tasks when planning an evaluation study is to describe the components of the project using a systems perspective. Relevant project components can be divided into four major parts: inputs, activities, results, and outcomes, as shown in Figure 1. The project itself exists in a larger, on-going criminal justice system and social system.

Inputs are the ingredients and elements received by the project from the environment in which the project exists. These include resources (funds for personnel, equipment, and so on), guidelines that constrain the operations of the project, and other rules or operating procedures (formal and informal).

FIGURE 1

SYSTEM AND PROJECT COMPONENTS



Activities include the operations of the project and any other organizational procedures, criteria, or rules that are developed within the project (rather than from outside the project). Anything that is done with the resources and other inputs by project personnel is included in the category called activities. Thus, the activities category includes those things produced entirely by project personnel, such as the provision of services, information, and so on.

Results are the initial consequences of project activities and include consequences which logically fall in between activities and outcomes, in a type of causal sequence. Results are distinguished from activities in that, although the project and its resources theoretically can produce activities, they cannot guarantee that the results will occur. A project, for example, can provide counseling services to its clients (an activity), but it cannot guarantee that the counseling will change the attitudes of the clients (a result). A project can engage in a public relations campaign to improve relationships with other agencies (an activity), but it cannot guarantee that relationships will improve (a result).

Outcomes are the broader-range, socially relevant, consequences of the project. In a sense, outcomes are those consequences of a project which, if positive, need no further social or political justification: They are an end in themselves. The provision of safety to the public (e.g., crime reduction or prevention) and the provision of justice are the two major outcomes from the criminal justice system. Thus,

project consequences closely related to these would be called outcomes and consequences that intervene between activities and the outcomes generally would be called results. The division between results and outcomes is quite arbitrary and, if one wished, the causal linkage of activities to results to outcomes could be placed in a whole series of boxes or categories rather than just those described here.

The feedback loop shown in the criminal justice system part of Figure 1 indicates that results and/or outcomes of a project can feed back into the operations of the project as additional inputs. A project responsible for soliciting clients to receive its services might have an educational campaign (an activity) to interest crime victims, for example, in the project services. A result might be that victims contact the program which, through the feedback loop, becomes an additional input. Information about the results and/or outcomes of the project also is channeled through the feedback loop into the project.

When an evaluator is beginning the planning phases for the study, it is useful to read the grant application and actually develop a systems-based, itemized description of the project. This could be done on a form such as that shown in Table 1. The specific inputs, activities, expected results, and expected outcomes (as presented in the grant application) are entered in the top part of the form. The lower section of the form is to be completed by the evaluator as s/he examines the logic of the project to determine whether critical ingredients are missing and what the essential intervening linkages are.

TABLE 1

SYSTEM DESCRIPTION OF THE PROJECT[1]

GRANT APPLICATION	INPUTS	ACTIVITIES	EXPECTED RESULTS	EXPECTED OUTCOMES
ASSUMPTIONS, IMPLIED ELEMENTS, LINKAGES				

[1]This form is a slight adaptation of the one currently used in the LEAA evaluation training course. In that course the form is called the "Method of Rationales."

The Logic of the Project

After the project description is completed (from the grant application and/or discussions with project personnel), the second task for the evaluator is to trace the logic or theory of the project. (This procedure is being called the "Method of Rationales" in LEAA's Evaluation Training Workshops. The use of the system description to trace the rationale of the project provides the name that LEAA has given to the chart.)

There are no rules or guidelines on how one goes about identifying the rationale of the project, but a few techniques are available that can assist the evaluator in the task.

One method of tracing the logic of the project is to start at the right-hand side of the system description and try to determine whether it is reasonable for each of the expected outcomes to occur and what the intervening events (between the activities and outcomes) are that must exist if the project is to accomplish the outcome. In doing this, the evaluator might think ahead to the types of discussions that would ensue if the project does not accomplish the outcome. The various reasons or rationales (or excuses) which might be given to explain why the project did not achieve the outcome or goal would provide a useful beginning point for identifying the assumptions that have been made as to why the activities ought to produce the desired outcomes. The variables identified as critical linkages, if eventually included in the evaluation, would provide information as to why the project did not work or, if the project is effective, the evaluator would be able to determine why it worked. The former information is very important in determining whether,

with changes, a project might be more effective in the future, whereas information on the linkages for a project that was effective might help in assessing whether the project is replicable in other situations or places.

Another technique for identifying the assumptions and intervening variables is to ascertain what types of behavior changes are needed if the project is to achieve the outcome (or the results). If certain behavior has to change, the evaluator should determine whether there appears to be sufficient incentive for this to occur. If it is questionable whether the incentives exist, data about the motivations or attitudes of persons whose behavior must change for the project to be successful would be a useful inclusion in the evaluation.

The logic of the project can be traced back from outcomes to results and from results to activities and from activities to inputs. In some instances, the evaluator may identify a complete causal chain linking all of these parts together and, if this is the linkage of major concern to decision makers, it would become the focal point of the evaluation.

Evaluators also will find it useful to start on the left-hand side of the systems diagram and assess the likely effect of each component of the system on the next. This approach assists in identifying critical inputs that may not have been provided for in the grant application, activities that are not included but which are necessary if certain results are to occur, and results that were not mentioned in the grant application but which have to occur if the outcome is to be achieved. In addition, this approach will help the evaluator identify unintended

consequences (positive and negative) of the project.

It should be emphasized that project directors, program developers, planners, and other persons in the system would find these same techniques useful in determining whether a project "makes sense." Problems in the rationale of the project identified by the evaluator or others could be called to the attention of those responsible for implementing the project and corrected even before the project begins.

Many evaluators will find it useful to actually fill in the lower portion of the system chart shown in Table 1 with a listing of the implicit assumptions, intervening variables that perhaps should be included in the evaluation, questions concerning what various aspects of the project actually mean, and so on. Thus, the system description form becomes a summary of the project and the potential independent, intervening, and dependent variables to be included in the evaluation.

The Project and Its Environment

As noted previously, a project is part of an on-going system and exists in the larger criminal justice and social environment. Many, but not all, evaluations involve a comparison of the project with the system as it existed without the project or a comparison of the project with some other alternative approach. In order to determine what the results and/or outcomes of the project are to be compared with, the evaluator should know how the project changed the system.

It is particularly important in terms of the development of cumulative knowledge concerning the effectiveness of alternative approaches for solving problems that the evaluator be aware of the theory upon

which the project is based and be aware of whether there are competing theories. If there are competing theories, the evaluator should attempt to compare the project with an alternative that is based on a different theory. Some projects may contain more than one component, with the different components representing different theoretical approaches to solving a problem. In this situation, it may be quite useful for the evaluator to compare one component of the project with another.

Selecting the Questions and Propositions

The procedures described above provide the evaluator with the knowledge and information needed to develop a complete set of recommendations concerning what questions the evaluation might attempt to answer. These, in turn, could be converted into propositions and hypotheses.

An evaluator who has prior knowledge of the types of issues or questions of concern to the eventual users of the evaluation results would not need to develop a complete set of all questions that could be answered, but evaluators often do not know exactly what the issues or concerns of all the potential users of the information are. The questions or propositions that the evaluator recommends to decision makers also are constrained by the level of funding available to conduct the evaluation, by problems of measurement, by the types of designs that are feasible, and by considerations for the confidentiality and privacy of data.

In some situations the evaluator does not formulate the questions to be answered, as this has already been done by project personnel, planners, or others in the system. Even so, the evaluator would be much better prepared to guide or modify those choices if s/he were aware

of the full set of questions that could be answered. Persons without technical training sometimes fail to note that certain questions could be answered and at other times assume it will be simple to obtain valid answers to questions when, in fact, it is impossible. Thus, the evaluator has a very important role in determining what questions or propositions will be incorporated in the evaluation.

The Type of Evaluation

The choice of performance measures (dependent variables) generally serves as the starting point for developing the evaluation questions or propositions. This choice can serve to determine what type of evaluation is being done. Within the LEAA evaluation training course the types of evaluation are identified by the point in the system where the final performance measure is taken. As shown in Figure 2, an impact assessment is an evaluation which attempts to link one or more outcomes in a causal fashion back to results, activities, and/or inputs. The assessment also may include examination of the effects of activities on results or of inputs on activities.

Process evaluation is defined as a study in which the final performance measures are results, rather than outcomes, and the results are linked to activities and/or inputs.

Monitoring in the LEAA training course is defined as a type of evaluation in which activities are linked (in a causal fashion) to inputs.

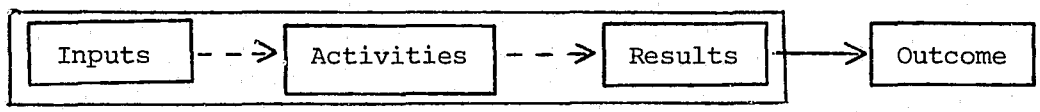
Dimensions of Performance

Regardless of the type of evaluation that is to be conducted, there are four major categories of performance dimensions that should be

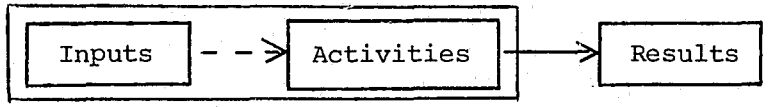
FIGURE 2

TYPES OF EVALUATION[1]

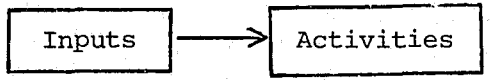
IMPACT ASSESSMENT



PROCESS EVALUATION



MONITORING



[1]These are the types and definitions of evaluation currently being used in the LEAA evaluation training course.

considered for potential inclusion in the evaluation. These are the quantity, quality (including equality), timeliness, and cost of a project. In an impact assessment where the evaluation is to focus on crime reduction, for example, most of the relevant questions about the effect of the project on crime are subsumed if one considers these dimensions, as illustrated below:

1. Quantity: How much crime was prevented?
2. Quality: Was the fear of crime reduced? Was serious or trivial crime reduced? For whom was crime reduced?
3. Timeliness: How long did it take for the effects to occur? How long will they last?
4. Cost: What did it cost to prevent how many dollars worth of crime?

These dimensions can be applied to outcomes (in impact assessments), to results (in process evaluations), and to activities (in monitoring).

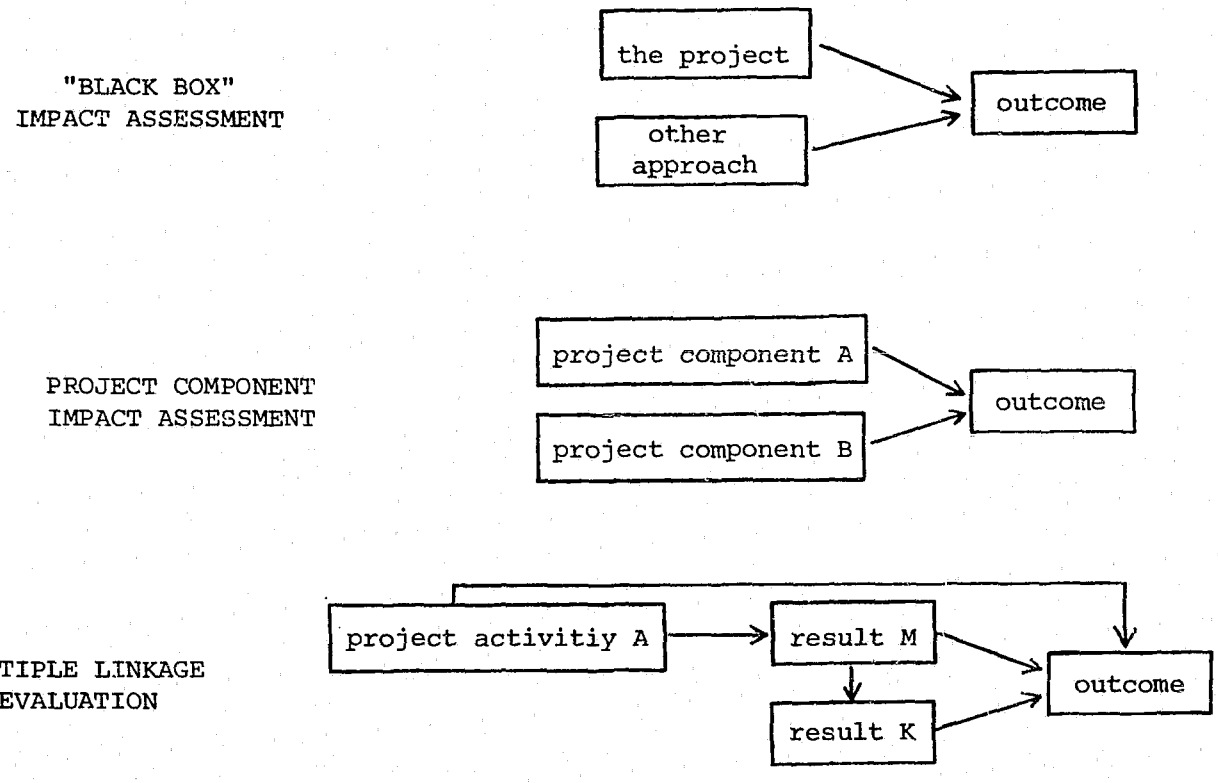
The Independent and Intervening Variables

The specific selection of independent and intervening variables depends very much on the project to be evaluated, the situation, the type of theory underlying the project, and so on. Nevertheless, there are three general strategies that evaluators could use.

1. "Black Box" Evaluations. A "black box" evaluation refers to one in which the entire project is compared, as a whole, with some alternative method of achieving the same or similar objectives and goals (see Figure 3). In an impact assessment, for example, the "black box" approach could involve a comparison of the recidivism rates of youths in the project with recidivism rates of youths handled through the

FIGURE 3

SELECTION OF INDEPENDENT AND INTERVENING VARIABLES[1]



[1]These are illustrations of the types of comparisons and linkages that can be used. Many combinations and extensions of these are possible. The same types of configurations can be used when results (instead of outcomes) are the dependent variables and when activities are the dependent variables.

"traditional" approach (whatever that might be). In a process evaluation that focuses on results, such as a reduction in school behavior problems of youths in a counseling program, the evaluation could compare project youths with similar persons handled through the traditional approach. Even in monitoring this strategy could be used. For example, the evaluator might compare the cost (per client) for providing counseling to rape victims by a project administered in the police department with the cost (per client) of a similar project operated as a non-profit.

2. Project Component Evaluations. As illustrated in Figure 3, one aspect or component of the project can be compared with another in terms of its effectiveness vis a vis outcomes, results, or activity levels. A project that provides counseling, tutoring, and social activities might be evaluated in terms of the effectiveness of each component (separately) or in relation to the best combination of these.

3. Multiple Linkage Evaluations. All evaluations involve establishing whether at least one linkage exists (the linkage between independent and dependent variables). A multiple linkage evaluation is one that attempts to establish whether one or more intervening linkages are critical to the success of the project (see Figure 3). A multiple linkage evaluation might seek to determine for whom or under what conditions certain effects occur. For example, an evaluation could focus on whether a crisis intervention program reduces detention days and other types of penetration into the juvenile justice system and whether these, in turn, reduce the recidivism rate. Alternatively, the crisis intervention counseling, per se, could have an effect on recidivism that is independent

of changes in detention time or other indicators of penetration into the system.

The list of questions, propositions, and hypotheses that could be included in the evaluation can be formulated after the evaluator identifies the potential performance measures, the candidates for independent and intervening variables, and the dimensions of performance that seem relevant. As noted previously, the funding level for the evaluation, measurement problems, protection of confidentiality and privacy, and problems in the evaluation design will have a strong influence on which questions actually can be answered. Thus, the evaluator needs not only to be aware of the potential questions, but must assess which of these can be answered within the constraints of budget, measurement, design, and confidentiality/privacy of data.

When the discussions or negotiations are held concerning the exact questions that will become the focal points of the evaluation, it would be very helpful if the evaluator could inform project personnel, planners, and other decision makers about what will be required in terms of funding, project operation, and so on in order to obtain valid answers to the questions. This would help prevent one of the more common problems that arises between evaluators and decision makers: The eventual inability of the evaluator to provide scientifically valid answers to questions that were of major interest. If fruitful discussions and negotiations are carried on between the evaluator and the eventual users of the information, a second common problem in evaluation might be avoided: The production of an evaluation report that did not examine the questions of interest to decision makers.

SECTION 3C

ALTERNATIVE APPROACHES FOR ESTABLISHING THE CRITERIA OF SUCCESS*

Abstract

Grant applications usually contain specific, quantitative statements of goals and objectives. These should be useful as a guide to the conduct of an evaluation, but they often are not, for reasons explained in this paper. The discussion focuses on alternative ways to specify the criteria of success and the role of evaluation in determining "success." In addition, a non-technical presentation is made on the use and interpretation of significance tests in determining project "success."

*This paper was written by Anne L. Schneider.

ALTERNATIVE APPROACHES FOR ESTABLISHING THE CRITERIA OF SUCCESS

Introduction

Determining whether a project, program, or policy is a "success" involves four quite distinct steps:

1. An identification, on the conceptual level, of the problems which the project was supposed to solve or ameliorate;

2. Selection of specific operational measures that are valid indicators of the concepts;

3. Selection of a particular "amount" of the problem that must be solved in order for the project to be deemed successful; and

4. Selection of a particular probability level in reference to the "amount" of the problem that the project is supposed to solve.

To illustrate: The problem that the project is supposed to ameliorate is the residential burglary rate; the indicators of this could be the reported number of residential burglaries divided by the number of residential units in the city; the "amount" of the problem that is supposed to be solved might be a five percent reduction in the rate of occurrence; and the probability that the five percent reduction could be produced by chance is 10 out of 100 (e.g., .10 significance level).

Although the procedure for defining the success of a project might appear to be straightforward and obvious, there are, in fact, several options that can be used at each point. The identification of the broad gauged problems which the project is supposed to solve generally is done within the grant application process. The evaluator is not particularly

involved in this, except to point out other positive or negative consequences of project activities that could be included in the evaluation and to assess the costs and difficulties of including these. The options (and difficulties) in selecting valid indicators of the concepts depend primarily upon what the goals are and, even though this is a very important step and is done mainly by the evaluator, it will not be discussed extensively here. Rather, the focus of this paper is to examine some of the choices concerning the "amount" of the problem that must be solved if the project is to be "successful" and to examine the choice of a significance level.

Magnitude of Project Impact

Perhaps the best method of determining the amount of a problem that a project must solve in order to be successful is to conduct a cost-benefit analysis. In this approach, one measures the total social costs of the project and its total social benefits. If the project can solve enough of a problem that its total social benefits are greater than its total social costs, it would be judged successful. Another method is to compare one alternative strategy with another (or with several others) in order to ascertain which is the most cost effective. With this technique, a project should be able to reduce the amount of a problem for a lower per unit cost than other alternatives in order to be judged more successful than the others.

Because of the difficulties in conducting these types of analyses,*

[1] See Section 3D, "Cost-Benefit and Cost-Effectiveness Analysis."

decision makers usually must determine the success of a project without the type of cost-benefit information that would permit perfectly rational decisions. Instead, one finds grant applications with quantitative measures of success that were apparently considered good enough so that, if the project were able to achieve those objectives and goals, the decision makers would continue funding it. Some evaluators object to these types of artificially established measures of success and would prefer that the project state its objectives in a way that makes it easier to convert them directly into an evaluation design.

Consider, for example, a statement that says the goal of the project is to reduce the recidivism rate of project clients by ten percent. Even though this might appear to be a very precise statement, an evaluator will recognize that its meaning is not at all clear. It could mean "to prevent ten percent of a group who otherwise would have recontact with the court from having a recontact." But with whom are the clients to be compared? It could also mean "to reduce the frequency of offenses, on the average, by ten percent compared with the frequency prior to the project." It could, of course, mean many other things as well.

Still other evaluators will claim that unless the project has quantitative goals, it cannot be evaluated. That claim is patently false.[1] An evaluator can measure the magnitude of project impact and report it, along with the probability that it was due to chance. Even if a project does not state what its general goals are, the evaluator can examine its activities in order to determine what types of social consequences could reasonably be expected from them. These, then, would constitute the dependent variables in the evaluation.

There are also some evaluators who permit the quantitative statements of goals or objectives to determine how the evaluation is conducted. Thus, an evaluation report might say that residential burglaries declined by five percent during the two years of project operation from the two years prior to project operation; therefore, the project achieved its objective. Even though the first part of the statement may be true, the second is not necessarily true at all. The task of evaluation is to establish a causal linkage between the project activities and the outcomes in such a way that the impact of the project on the problem can be ascertained as independent of other factors and as independent of chance.

Although there is general agreement that the causal linkages have to be established, there is no consensus concerning the most appropriate method of stating quantitative goals and objectives for the project or of determining project success.

Some of the alternatives used in evaluation for defining and testing project "success" are described below.

1. The evaluator can simply ignore the quantitative statements in the project grant application and test to determine if the project impact is significantly different from zero. If this approach is used, the evaluator should at least include data showing the best estimate of the actual magnitude of effect. The decision makers ultimately responsible for deciding whether the project is a success or not will have the information that they can compare with the "promises" made in the grant application and can decide whether the impact is large enough to suit them.

2. The evaluator could test whether the project impact is significantly different from zero and could also conduct a test of significance

to see if the impact is significantly less than the quantitative objective stated in the grant application. In other words, the null hypothesis (for the second test) is that the project reduced crime, for example, by five percent or more. If this hypothesis is rejected, then the project "failed" at least in terms of its stated objectives. In some evaluations, of course, it is possible that the evaluator will not be able to reject the null hypothesis in either test. That is, s/he cannot reject either the possibility that the crime rate was the same or the hypothesis that crime was reduced by five percent.

3. Confidence intervals are another possible option for the evaluator. In this procedure, it does not matter whether the project stated quantitative objectives or not. The evaluator calculates the confidence interval (.05 or .01 or whatever) around the figure representing the magnitude of impact. For example, the observed reduction in crime attributable to the project might be 15 percent, with the lower bounds of the confidence interval being 10 percent and the upper bounds being 20 percent. The 15 percent is <u>still</u> the best estimate, but the lower and upper limits provide additional information to the decision makers.

4. Some evaluators restate project objectives so that the objectives or goals are to achieve statistically significant changes (or differences). This approach has the advantage of bringing about congruence and agreement between evaluator and decision makers before the evaluation is conducted and could improve the likelihood that decisions will be based at least in part on the evaluation results. On the other hand, this approach is rather inflexible because the size of the sample has a strong influence on the amount of difference needed to achieve statistical significance. Thus, a project for which the evaluator collected recidivism

data on 1,000 treatment youths and 1,000 control youths will find it very easy to achieve statistical significance even though their impact could be quite trivial (a 2 percent difference--i.e., 8 percent versus 6 percent--is significant at the .05 level with this many cases). A project with 50 treatment cases and 50 control cases would find that even a substantial difference of 15 percent (such as 50 percent versus 35 percent) was not statistically significant at the .05 level.

5. Another option available to evaluators when the project has stated quantitative objectives but not indicated the method of comparison is to restate the objectives incorporating both the sample size, method of comparison, and the quantitative goal of the project. For example, if the project goal is to achieve a 50 percent decrease in recidivism (such as from 20 percent down to 10 percent), the evaluator could restate the objective in the following way: "The goal is that the recidivism rate of 200 project youths, 12 months after entry into the project, will be 50 percent less than that of 200 control group youths, 12 months after their entry into the system (i.e., the control group will have a 20 percent recidivism rate and the experimental group will go from the 20 percent rate down to a 10 percent rate)." With a sample size of 200, the difference between 20 percent and 10 percent is statistically significant beyond the .05 level. This approach can be used if the evaluator has a good estimate of the size of the percentage and can draw samples large enough to show significance if the desired percentage difference is achieved.

An argument could be made that it is not the evaluator's job to determine whether a project was "successful" or not and, therefore, the

evaluator does not need to be concerned with how the project personnel or other decision makers--such as those in the planning agency--would define "success" for any particular project. This argument is based on the idea that an evaluator, as a social scientist with considerable knowledge of the criminal justice system, is perfectly capable of examining the project activities and, from these, projecting what the potential effects of the project are on its clients, other parts of the criminal justice system, and the community. Furthermore, the evaluator has the technical expertise to establish a valid method of comparison. In the evaluation report (so the argument would continue) the evaluator reports the magnitude of impact the project had (positive and negative) on other parts of the system, clients, and/or community and the probability of whether these impacts were due to chance rather than to the project. At that point, it is up to the decision makers to decide whether the project was "successful" or not in relation to the impacts included in the evaluation.

On the other hand, persons concerned with the utilization of evaluation results in decision making would argue that the evaluator and decision makers (such as project directors, planners, elected officials, and so on) should discuss and negotiate the criteria of success, including the specific methods of comparison and the actual "amount" of impact that will be considered sufficient. Through this process, the persons responsible for making recommendations or decisions might become more aware of the value of evaluation and what is needed to obtain valid answers. Evaluators would become more sensitive to the wide range of questions that are of interest to the eventual users of the evaluation

results and the range of performance measures that could be included.

The Level of Statistical Significance

The purpose of this discussion is not to present a technical discourse on tests of statistical significance, but instead it is to provide some guidance on the use and interpretation of significance tests within the context of evaluation research.

A test of significance is conducted in order to estimate the probability that tan apparent impact of the project on the dependent variable could have occurred by chance and, therefore, the project should not be credited with the outcomes or results even though they occurred. In laboratory experimental research it is an accepted practice to establish the significance level at .001 or .01 or .05 before conducting the experiment and to adhere rigorously to the principle that one must assume "no effect" unless the probability of a chance occurrence is quite low, such as one in a thousand, one in a hundred, or five in a hundred.

Applied research differs in several fundamental ways from basic research and the use of tests of significance should be somewhat different.

First, it must be recognized that whenever one observes a change in the dependent variable (the criterion of success) which might be attributable to the project, there almost always are several alternative explanations for why the change occurred. Similarly, in a comparison design, if one observes a difference between the experimental and comparison groups there may be several potential reasons--other than the project--that could account for it. One possible reason for

a change or a difference is a purely chance occurrence. Tests of statistical significance provide information on the likelihood that chance is the alternative explanation, but these tests do not provide any information at all on whether other alternative explanations (i.e., threats to validity) could have produced the observed difference or change. Thus, in evaluation research tests of significance always should be conducted, but it must not be assumed that a statistically significant effect shows that the project was the factor influencing the change in the level of problem unless other confounding factors can be ruled out.

Conversely, external factors not related to the project could serve to minimize or hide completely the impact of a project even though it had a substantial effect on the outcome of interest. A project might, for example, simultaneously increase the reporting of rape cases and reduce the incidence of rape. If the rate of this crime (as officially reported to authorities) is used as the measure of success, it might increase due to the increased reporting or it could decrease because of a reduction in the true (reported and unreported) frequency. Thus, one must not conclude that the failure to achieve a significant difference shows that the project was ineffective unless it can be established that external factors did not hide or mask the effects.

Third, it was noted earlier that the size of the samples has a considerable effect on the amount of change or difference that must occur if statistically significant differences are to be obtained. An evaluator could find himself or herself in the rather embarrassing situation of having such small sample sizes and low base rates on the dependent variable that it is impossible to establish statistical

significance for the project even if it reduces the problem to zero. It is conceivable, for example, that the recidivism rate or crime rate within a control group or area to be so low and the sample sizes so small that a reduction to zero in the experimental group would not be statistically significant.

Fourth, evaluators should recognize that decision makers often have to choose from among many alternative ways of allocating funds and have very little information on the probability that any of the alternatives will have an impact on the problem. They may not require that the chances be 5 out of 100 before they are willing to take a chance that a particular project is a better option than the other approaches. Evaluation reports should include the exact probability level from the significance test rather than just a statement of significance or non-significance at the .05 level (or .01 level or whatever). Decision makers might decide that a project which was evaluated using small samples and which indicated an impact with the associated probability that it would occur by chance 20 times out of 100 to be a better risk than any other option available to them.

Fifth, tests of significance in applied research often are conservative (underestimates) of the magnitude of the impact of the project--providing that the alternative explanations for any observed differences or changes can be ruled out. The tests are conservative due to the low reliability of the data generally used in criminal justice research.

Tests of significance are conservative in another sense. It is normally the case in criminal justice that only the newer, more innovative projects are evaluated at all. These approaches are required

demonstrate effectiveness beyond reasonable doubt, whereas other approaches that are more a part of the established system do not have to demonstrate effectiveness at all. Most projects are evaluated in comparison with whatever existed in the past or with a concurrent comparison group or area that is receiving the "usual" treatment or procedure. It is the new approach that usually is required to be significantly better than the old. In some situations, it would be just as reasonable to require that the traditional approach be significantly better. In other situations it would be quite reasonable only to require that the new approach not be any worse than the old. Examples of this include the deinstitutionalization of status offender (DSO) projects. Most of these projects remove status offenders from the jurisdiction of the juvenile court and may be less expensive and more humane in the sense that status offenders cannot be detained or sent to state institutions. Nevertheless, many of these projects stated as a goal that the recidivism of the DSO status offenders would be less. One could argue that it is not fair or reasonable to require a statistically significant recidivism reduction in comparison with the traditional approaches in order to establish the success of the deinstitutionalization projects.

FOOTNOTES

1. It is the case that some projects cannot be evaluated, but the source of the problem is not the lack of quantitative goals. Rather, problems stem from such things as weak designs, unreliable data, or no data at all.

SECTION 3D

COST BENEFIT AND COST EFFECTIVENESS EVALUATIONS*

Abstract

Cost benefit analysis requires measurement on a common unit of value of the full social costs and full social benefits of a program. Cost effectiveness analysis is less complex in that it requires comparable measurement of the costs of two or more strategies and measures of effectiveness vis a vis one (or more) common objective. The requirements of cost benefit and cost effectiveness analysis are explained and an example of cost effectiveness analysis is presented.

*These materials were presented by Anne L. Schneider at a special forum of the Model Evaluation Program.

COST BENEFIT AND COST EFFECTIVENESS EVALUATIONS

Introduction

The purpose of this paper is to explain, on a conceptual level, the differences between cost benefit analysis and cost effectiveness analysis and to explain how one does cost effectiveness analysis.

Cost Benefit Analysis

Cost benefit analysis is the best, most comprehensive type of evaluation that one can conduct on any type of public policy or program.[1] It unfortunately is also far more difficult and complex than any other type of evaluation--so much so that it is often impossible to conduct cost benefit analysis for social programs.

The key characteristics of cost benefit analysis are described below:

1. One must first measure the total social benefits of a particular policy. Social benefits are defined as the sum total of all benefits directly attributable to the policy from the perspective of the society as a whole. Benefits mean "wants" and include all things valued by the society.

Direct benefits of the criminal justice system would include benefits such as safety from crime, justice, and various types of emergency services available to the society.

Indirect benefits include goods or services that are "external" to the direct benefits. For example, the criminal justice system in Seattle produces direct benefits for Seattle. If Seattle develops a highly effective program for preventing juvenile delinquency and if a proportion

of these youths move to Spokane, then the subsequent reduction in juvenile crime in Spokane is an indirect benefit of the Seattle program.

2. After measuring social benefits, one must then measure the total social costs of the policy that is being considered. Social costs are defined as the "wants" or "benefits" that could be produced for the society as a whole if the resources used for the policy under consideration were, instead, used for something else or were left to the private sector (that is, the money was never collected in taxes). The social costs also are called opportunity costs. If the resources to be used for the policy are used for it, then one must know what opportunities are foregone by this choice of expenditure. Thus, the first component of social costs is the value of the resources being used for the policy, measured in terms of alternative uses. Most commonly the value of the resources is assumed equal to the dollar value if the funds for the program had been left to individuals and never collected in taxes.

A second component of social costs is the direct or indirect negative result--if any--of the policy. For example, if Seattle develops a very effective enforcement program against drug use which displaces drug pushers and users to other parts of the state, then the transplanting of the problem is an indirect cost to the other areas of the Seattle program. If persons in Seattle are conducting the cost benefit analysis, they could define the society as "Seattle" and ignore these displacements; but if the state is conducting the analysis, then these need to be counted as costs of the Seattle program to another area.

Another example of indirect costs are costs incurred by clients or citizens. If Spokane develops a program for youths that requires

CONTINUED

3 OF 7

parents to incur costs--travel time, appearing in courts, etc.--then these are additional indirect costs of the program.

3. The costs and the benefits now must be converted to a common scale. For example, if some of the costs are measured in resources and dollars, then one cannot measure benefits in terms of recidivism reduction, or improvement in community support for the criminal justice system, and so on, because these measures are not comparable to dollars. In practice, all the costs and benefits usually are measured in dollars. The exact amount of the "benefit" attributable to the program has to be known or estimated. For example, if juvenile crime declines as a result of HB 371 and if someone is successful in figuring out the dollar value of reducing juvenile crime by a particular amount, then one must determine exactly how much of the decline is attributable to the new programs and procedures introduced by HB 371. The amount that is attributable to something else (for which costs have not been measured) cannot be included as a benefit from the bill. The proportional amount is all that should be attributed to the program because only the costs of the program have been measured. The costs of other factors that may also have reduced juvenile crime have not been measured. Serious errors result if benefits are attributed to a program when, in fact, they are the result of something else and the costs are not included in the analysis.

4. Having achieved all of the above, one can now calculate the cost benefit ratio. If the total social costs are greater than the total social benefits, then the government should not be spending the money on the program since this results in a net loss to society. If the total social benefits are greater than the total social costs, then

this is a proper expenditure of funds.

5. There is a great deal more to cost benefit analysis than this brief overview. One needs to calculate the cost benefit ratio for this year, next year, 10 years in the future, and so on, to determine whether or not the government should invest taxpayers' money in the program. There is no point discussing techniques of extending the analysis into the future, because, with virtually all criminal justice programs, a true cost benefit analysis is impossible.

Cost Effectiveness Analysis

Cost effectiveness analysis is simpler and is something that can be done in relation to social programs.[2] Cost effectiveness analysis is not something new--it basically involves adding a "cost" component to the usual types of effectiveness-oriented evaluations. The basic procedures are described below:

1. One must select one or more measures of effectiveness. These can and often do include social benefits such as those used in cost benefit analysis, but the value of the outcomes do not have to be converted to dollars. Thus, one could measure recidivism reduction, crime rates, "justice" or perceptions of it, community attitudes, and so on, using different scales for each of these.

2. The investigator must specify more than one method of achieving the goals. That is, a new program is compared with an old one; two new programs are compared to each other; or two or more variants of the same program are compared, and so on. Cost benefit analysis can (theoretically) be conducted on just one program. With cost effectiveness at

least two alternative ways of trying to accomplish the goals are needed.

3. The cost of each program must be measured. There is more flexibility concerning how this is done than with cost benefit analysis, but one must be careful to insure that the method of measuring cost for program A is comparable to that used for program B. For example, the budget (personnel, supplies, etc.) for each program could be used, but if one of the programs required capital expenditures and the other did not, in the long run the capital expenditures of the first program would become less and less important. Measurement of the cost can get rather complex, but the basis procedure is to figure out what the components are of each program and then add up the costs of personnel, supplies, value of space used, etc.

4. Next the investigator needs to measure the effectiveness of the program in relation to the measures selected for it. This is done in the manner usually used in evaluation. The amount of recidivism, for example, from one program is compared with the alternatives. As with cost benefit analysis, one must have a precise estimate of how much of the difference in recidivism (or reduction, if using pre-post measures) clearly is attributable to the program. In addition, one may find it useful to put a confidence interval around the number. Thus, if program A's recidivism level is .20 and program B's is .30, the difference of .10 is statistically significant and the confidence interval for the difference might be .08 to .12.

The data in Table 1 illustrate a cost effectiveness analysis of two programs. Each program begins with 100 clients. Program A has a recidivism rate of 25 percent, compared with a recidivism rate of

TABLE 1

AN ILLUSTRATION OF COST EFFECTIVENESS ANALYSIS[1]

Data & Calculations	Program A	Program B
DATA		
(a) number of clients	100	100
(b) recidivism rate (subsequent convictions)	25%	20%
(c) number of clients NOT recidivating	75	80
(d) cost of project	$7,500	$16,000
CALCULATIONS		
1. cost of preventing a subsequent offense by non-recidivators[2]	$\frac{\$7,500}{75} = \$100$	$\frac{\$16,000}{80} = \$200$
2. marginal additional cost of program B to prevent 5 more offenses than program A		$16,000 - 7,500 = \$8,500$

[1] It is assumed that the differences in recidivism are due entirely to the differences in program operation and not to any other factors.

[2] This calculation assumes, in a sense, that everyone in each program would have recidivated if the program had not existed.

20 percent for program B. The number of clients not recidivating is 75 for program A and 80 for program B. The cost of program A is $7,500, compared with $16,000 for program B. The first calculation shows the cost of preventing each subsequent offense. If it is assumed that 75 subsequent offenses are prevented by program A and 80 by program B, the per unit cost for program A is $100, while the per unit cost for program B is $200.

Marginal comparisons also should be made. Program B prevented five subsequent offenses more than program A for a cost of $8,500 more. Thus, it costs $8,500 more to prevent five more subsequent offenses.

In this example program B is more effective in reducing recidivism than program A (rate of 20 and 25 percent, respectively). But the critical question is whether it is worth $8,500 more to "prevent" five more persons, per 100, from recidivating.

Cost effectiveness analysis normally involves a comparison of two alternatives in relation to one objective at a time. It can, however, be extended to incorporate multiple objectives through the use of subjective valuation of alternative goals. The procedures are to have decision makers (or the public) place a subjective estimate of the dollar value on the achievement of each unit of each goal. For example, consider an evaluation that includes measures of recidivism and community support for the justice system. The latter might be measured on a five-point scale ranging from a high level of support to a very low level of support. Decision makers would need to place a dollar value of preventing each subsequent offense and a dollar value on community support. The latter might be accomplished by specifying the dollar value of having

100 percent of the community in the "high support" category, compared with having 90 percent in that category, 80 percent, and on down to zero.

After measuring the effectiveness of program A and program B in relation to each goal (recidivism reduction and community support), a final cost effectiveness score for each program can be calculated. Generally, however, one would simply show the cost and effectiveness scores for each program in relation to each objective and let the decision makers judge the overall merit of each.

It should be emphasized that both cost effectiveness and cost benefit evaluations are more difficult and more complex than normal outcome oriented evaluations. Neither cost effectiveness nor cost benefit avoid the "problem" of having to develop scientifically reliable methods of measuring effectiveness. Thus, neither approach can be used in the absence of some type of scientifically reliable methodology, such as experimental designs, interrupted time series, or comparison group designs.

FOOTNOTES

[1]A useful article on the cost-benefit analysis is Jerome Rothenberg, "Cost-Benefit Analysis: A Methodological Exposition," in Marcia Guttentag and Elmer L. Struening (eds.), Handbook of Evaluation Research Volume 2, (Sage Publications, 1975). Also see, Robert Dorfman (ed.), Measuring Benefits of Government Investments (The Brookings Institution, 1965).

[2]See Henry M. Levin, "Cost-Effectiveness Analysis in Evaluation Research," in Guttentag and Struening, Op cit.

SECTION 3E

THE ROLE OF EVALUATION IN RATIONAL & BARGAINING DECISION MAKING PROCESSES*

Abstract

The purpose of the paper is to explain how evaluation can be used in both rational and bargaining decision-making situations. In rational decision making, the evaluation findings are used to estimate the magnitude and probability of project effects for two or more strategies in order to choose between the strategies (or projects). In bargaining processes, evaluation findings are used to determine whether a project lived up to the "promises" in the grant application. In either instance, the evaluation must establish the causal relationship between the observed outcome and the project(s) being evaluated.

*A revision of a section in the "Final Report: An Assessment of Factors that Constrain and Facilitate the Use of Evaluation Information by LEAA Planners and Decision Makers in the State of Washington" by Anne L. Schneider and Peter R. Schneider, February, 1977.

THE ROLE OF EVALUATION
IN RATIONAL AND BARGAINING DECISION MAKING PROCESSES

One way of viewing a rational decision making process, within the contest of a democratic political system, is that goals (or problems) are selected in order to maximize public preferences concerning the actions of government, but the means of achieving the goals are determined on the basis of technical, scientific information--such as information from evaluations--concerning the effectiveness and cost of alternative strategies.

The phrase "rational decision making" refers to a particular method of arriving at a decision. In brief, it means that the decision maker selects the policy alternative which s/he believes has the highest probability of maximizing a benefits-cost ratio. In the contest of governmental decision making, in a democratic society, a constraint can be added to the definition. The public policy maker is said to engage in rational decision making if s/he supports the policy position that will have the greatest probability of maximizing the benefit-cost ratio to the public.

The citizens select the goals, and the scientific information is used to calculate the "best" means of achieving the goals. The distinction between the goals and the means of achieving them presumes that citizens either have no preferences about alternative methods of solving a problem or that their preferences would be the same as those of a rational decision

maker if they had the same information. This presumption probably is correct for the "general public" who will not suffer serious costs or obtain substantial rewards (in the short run) from any particular action that is being considered. The assumption obviously is incorrect for other persons and groups, since the choice of a particular strategy or project will distribute substantial benefits to some persons rather than to others. Virtually all decisions concerning the allocation of scarce funds result in some persons or groups "winning" a great deal and others "losing" a great deal.

If the prevailing philosophy of a political system or government agency is that public or group preferences should be relied upon to select goals (or to identify problems), but scientific information should be used to select strategies and specific projects to achieve those goals, then the role of evaluation in decision making is quite clear. Evaluations and other types of research should be relied upon to select the strategies and specific projects which have the greatest probability of achieving the goals for the least cost.

Another, perhaps more realistic, view of the political decision making process is that it inherently involves bargaining by identifiable groups with public agencies, not only in relation to goals but also in terms of strategies and specific projects. The bargaining is undertaken so that the group can maximize the allocation of funds to its membership through the funding of specific projects. In the specific context of LEAA funding, persons representing areas, clients, and government agencies other than LEAA bargain with those who distribute LEAA funds.

Decisions produced through these types of bargaining procedures could represent compromises and tradeoffs among the groups involved (or promises of benefits or punishments to the funding agency) without any necessary attention to the interests of the "general" public. On the other hand, the bargaining process can be constrained and structured by the funding agency through the use of guidelines that require proposals for funding to address public interest goals. In a bargaining system, groups or agencies that wish to obtain funds make "promises" to the funding agency concerning their performance and funding is based on these "promises".

Information produced from evaluations is just as critical in a political system where bargaining extends to the choice among means and projects as it is in the system described previously. Impact or process evaluations should be conducted to determine which of the competing groups (and the strategies they used) fulfilled the promises made about short or long-term public interest goals and the cost of achieving them. Monitoring would be used to determine when the operating procedures and activity levels are in line with the specific promises that were made by the groups when they struck their bargain with the agency.

In this type of system it is critical that at the second round of decision making and funding allocations the groups and agencies that did not live up to their promises would be penalized and those that did would be rewarded--or at least, given the vagaries of LEAA funding would be continued for another year. Through an incremental (and perhaps slow) process, information from evaluations should serve to winnow out the groups that cannot achieve social goals with a sufficient degree of efficiency and produce information on the types of activities, organizational patterns,

rules and strategies, that characterize the groups which consistently produce cost effective solutions to public problems.

The kind of evaluation used in the bargaining decision-making system does not differ much from those used in a more rational process. Its purpose is to establish a causal linkage between the project and the desired (i.e., promised) outcomes or results. The major difference is that evaluations in the rational process would be more likely to make explicit tests of two (or more) alternatives for achieving the outcomes or results.

There is a considerable risk in a bargaining model of decision making that the funding agency will not have sufficient clout to impose judgments based upon the results of an evaluation process. On the other hand, a funding agency that operates in a bargaining political context probably would find it even harder to impose requirements that funds be allocated strictly on the basis of the probability that a particular strategy or project would achieve an impact if this meant total disregard for the political clout of the bargaining groups. Funding agencies that operate within a political system which values public and group preferences in establishment of goals but excludes them in the selection of specific projects will not encounter the same types of difficulties in achieving a sufficient degree of reliance on evaluations.

The multi-layered aspect of decision making--especially when federal funds are involved--can introduce considerable confusion into the entire process because it may be difficult to determine the point within the system where effective decision are made. Evaluations need to be available and used at the point where the bargains and promises that determine actual funding allocations are made, even if this is not the final (official)

decision point. In a similar way, evaluations are needed at the point where strategies and projects are selected in the more rationalistic type of political system.

SECTION 4

APPLICATIONS OF PROBLEM SOLVING TECHNIQUES IN CRIMINAL JUSTICE EVALUATION

Abstract

The nine evaluation reports or excerpts from evaluation reports in this section demonstrate the use of one or more problem solving technique of interest and value to evaluators. Introductory comments which identify or expand upon the particular techniques of major interest have been prepared and precede each of the reports or excerpts.

SECTION 4.A.

Evaluation Report
City of Seattle
HIDDEN CAMERAS PROJECT*

By

City of Seattle
Office of Policy Planning
Law and Justice Planning Office

Lawrence G. Gunn
Director

Kenneth E. Mathews, Jr., Ph.D.
Senior Researcher/Evaluator

Antoinette Hood
Research and Evaluation Aide

January, 1978

*This is the full text of the Hidden Cameras Project Evaluation.

INTRODUCTORY COMMENTS ON:

"HIDDEN CAMERAS PROJECT EVALUATION REPORT"

This evaluation report by Kenneth Mathews and Antoinette Hood is an exemplary one in several ways.

One of the most instructive aspects of the evaluation involves the techniques used by the evaluators to achieve a field experimental design with random assignment. It should be noted that the evaluators were involved in the initial development of the program along with the project director and others in the planning office. The pre-project analysis conducted by the evaluators identified the characteristics of high risk establishments. Since the number of cameras was less than the number of establishments, it was possible to randomly select businesses for the cameras without a denial of service to anyone.

The assessment of project effectiveness was conducted by first establishing a causal relationship between the project and several indicators of system level performance and then estimating the cost-effectiveness of the hidden cameras. This section of the evaluation illustrates the complexity of measuring the costs of the project and the savings to the system, but it also demonstrates that cost-effectiveness evaluations can be conducted in criminal justice.

Although random assignment was used to examine project impact on arrests and convictions, the impact on commercial robbery rates of the entire city could not be assessed accurately just by comparing experimental and control sites. Thus, the evaluators used quasi-experimental designs for this part of the analysis. Their use of multiple time series is especially instructive in that different types of alternative explanations

for the findings were examined with the use of several different comparison groups.

Readers also should note the procedures used by the evaluators to examine many alternative explanations for almost every finding in the study. Initial findings, either of effectiveness or ineffectiveness, often were tested again, using a different design or different performance measure. The consistency in the findings from the multiple designs and multiple indicators of success permitted the authors to draw strong, unequivocal conclusions.

Hidden Cameras Project
Evaluation Report

Between 1966 and 1975, robbery in Seattle increased from 650 reported cases to 2,103, or by 224 percent within ten years.[1] When examined on a per capita basis (number of robberies per 1,000 City residents), the increase is even larger, 252 percent, or from 1.19 reported robberies per 1,000 residents in 1966 to 4.18 per 1,000 in 1975.

The large increase in robberies, combined with both the high potential for and actual occurrence of physical injury and financial loss to victims, resulted in robbery being chosen as a priority crime within the City of Seattle. Analyses of Seattle robbery data collected between 1972 and 1975[2] suggested that commercial robbery (which comprised 22 percent of all robberies in 1975) was potentially a good target crime for the arrest and prosecution of offenders in an attempt to reduce the overall rate of robbery. Specific factors leading to this conclusion were first, that the distribution of commercial robberies is concentrated within relatively few "high-risk" types of businesses; in addition, within the high-risk types of establishments, there were readily identifiable individual businesses with more potential for robbery than other businesses within the same high-risk group. Second, the recidivism rate among commercial robbery offenders was judged to be much higher than for other offender groups. This implied that the arrest and conviction of commercial robbers would result in the relatively permanent reduction of the total number of robbers, each of whom was believed to be responsible for multiple robberies.

Funding and Organizational Placement and Staffing

Initial project funding (90 percent LEAA, 5 percent State and 5 percent City) was $50,000 for the 18-month period of December 1, 1975, through May 30, 1977. The project budget was as follows:

$11,414	Personnel compensation (installation and service technician)
28,700	Equipment (including 75 surveillance camera units and miscellaneous project equipment)
9,886	Supplies and operating expenses (primarily film, camera repair and maintenance, and 5 percent administrative fee)
$50,000	Total grant costs

The project operated within the Crimes Against Persons Section of the Criminal Investigations Division of the Seattle Police Department. A police officer

---

[1] City of Seattle Criminal Justice Plan - 1977, Law and Justice Planning Office, Seattle, 1976, p. 28.

[2] Ibid, pp. 303-334

(Officer J. D. Nicholson) served as project director, to insure prompt equipment selection, purchase and installation, to assist in relevant target selection and data collection techniques, to manage day-to-day operations and to supervise the grant-funded technician and camera installations.

Project Intent, Goal and Objectives

The Hidden Cameras project was to place disguised cameras within potential robbery targets, and to encourage victims to record any robbery that occurred by activating the police-owned camera unit. The manner of camera activation was designed to minimize the chance of offenders knowing the camera had been started. The photographs were to be used as evidence in naming, apprehending and prosecuting offenders.[3]

The goal of the project is to increase the apprehension of robbery offenders and test the feasibility, efficiency and effectiveness of the portable, police-owned surveillance camera strategy.

The specific project objectives are as follows:

1. To increase significantly[4] robbery clearances by arrest for those businesses in which hidden cameras are installed as compared to other comparable businesses.

2. To increase significantly the proportion of convictions for commercial robberies in which photographs are taken as compared with those commercial robberies not involving hidden cameras.

3. To reduce significantly the cost of processing robbery cases from initial police response through investigation and prosecution and final court disposition for those cases involving hidden camera photographs as compared with other commercial robbery cases.

4. To reduce significantly the incidence of commercial robbery in the City of Seattle, as compared to other comparable jurisdictions.

5. To accomplish project objectives without significantly increasing the risk of injury to victims, bystanders, police and offenders.

---

[3]Actual site selection procedures, camera disguise and techniques used to take pictures will not be discussed in detail, in order to preserve the covert nature of the project. Individuals or agencies interested in operational details should contact Officer J. D. Nicholson, Project Director, Hidden Cameras Project, Seattle Police Department.

[4]Unless otherwise stated, statistical significance was $\alpha = 0.05$, or the chance of finding differences as large as were obtained would occur five times out of 100 due to sampling, if there were no real differences among the groups that were studied.

Project Operation

During the 10.5 months following first camera placement (mid-June 1976, through April, 1977), the following data on project operation and camera reliability were noted by project personnel.

Robbery Photographs: With 75 cameras in commercial sites, 38 robberies occurred, of which 32 (84 percent) yielded photographs of the crime. In addition, five offenses other than robbery (forged prescription, shoplift, burglary and till tap) were also photographed. This represents a rate of one robbery per camera site every 1.7 years.

Of the six robberies not photographed, three (50 percent) were because of a prior accidental activation that had not been detected and reset; two (33 percent) were due to activation failures not the fault of equipment or victim; and one was due to camera/equipment failure.

Photograph quality was sufficient for offender identification in all cases except one case in which there was a shutter speed malfunction.

Camera Failures and Service Requirement: During the period of time covered by the evaluation, 75 cameras were placed in sites for approximately 315 days each. This gives a total of 23,625 camera-days of potential operation (75 x 315). However, during this time, 26 camera-days were lost because of service/repair problems.[5] Assuming that all failures were immediately detected, operational camera-days would be 23,625 minus 26, or 23,599. This represents 99.89 percent of all the potential camera-days, showing very little time lost for service. To maintain a 99.0 percent coverage rate, failures would need to be detected within an average of 16.2 days.[6]

False Activation Rates: Within the 110 experimental locations in which the 75 cameras were placed at some time during the period (see Evaluation Design Section below), there were 315 false activations, or an average of 2.86 per camera. This represented a requirement for a site visit to reset a camera every 1.47 days.

While 22 sites (20 percent) never had a false activation, 88 (80 percent) had at least one. Within the 88 sites, the frequency of false activations was as follows:

---

[5]The vast majority of service requirements (12, or 92 percent) were due to accidental activation which resulted in automatic film advance mechanisms jamming. The remaining service requirement was for a faulty signal light.

[6]This was calculated by taking the total camera-days minus 99 percent minus service days required, or 23,625 minus 23,388.75 minus 26, which equals 210.25; this figure was then divided by the number of failures, which was 13.

Number of False Activations	Number of Sites
1	34
2	20
3	19
4	3
5	5
6	3
7	3
8	1

Because of numerous false activations and other problems, four cameras were removed from sites.

Evaluation of Objectives

Evaluation Design: Before project operation began, the project director and Seattle Law and Justice Planning Office research and evaluation personnel collected available information on all commercial robberies occurring during a preceding 18-month period. The robberies were classified by type of business and then summed to determine the number of robberies committed against each type of business. Bureau of Census data on the number of such businesses in Seattle were then used to estimate the type of business with the highest risk rate. Within these identified groups, those specific businesses with past robberies were chosen as the most likely to be robbed again in the future.

Based upon these data and other information, 150 commercial sites were identified as being the most likely places to be robbed in Seattle. These sites were then randomly assigned to either experimental (receive temporary camera placement) or control conditions.

Following site selection, store owners and managers were approached and asked if they wished to take part in the study. Cooperation with the project was very good in that only one possible site refused camera placements. As stores were involved, some sites were dropped from the study because of physical features or conditions making it impossible to place cameras. In these instances, a new store was identified as a probable robbery site, added to the control group and then a replacement experimental site was randomly selected from the control group.

If, in the judgment of the project director, the number of false activations in a store exceeded an acceptable number within a time period, the store was dropped from the study and replaced.

Approximately three months following initial camera installation, half of the cameras were randomly selected to be moved from their sites and then randomly assigned to control sites. At this point, the old experimental sites were designated as control locations, while the old control locations became experimental locations.

While it had been planned to continue to move half of the sites every two to three months, this did not occur after the first movement of cameras.

Objective 1: To increase significantly robbery clearances by arrest for those businesses in which hidden cameras are installed as compared to other comparable businesses.

Robberies may be cleared in one of two main ways. Either a crime is cleared through the arrest of the suspect, or it may be cleared "exceptional." Exceptional clearances involve instances in which the identity of the offender is known, but the offender is unavailable for arrest (dead, in prison, etc.), or the victim refuses to prosecute (the latter being relatively rare for robbery).

To examine the project effect on clearance rates, data were collected for every offense occurring at each of the 150 study sites. A member of the SPD Crimes Against Persons Section coordinated collection of data and forwarded completed data collection forms for each offense to the Law and Justice Planning Office for data analysis.

At the 150 sites, 100 offenses were reported from mid-June, 1976, to April 27, 1977. Ninety-four of these offenses were for robbery, and six were for other crimes, of which five were photographed (till tap, shoplifting, forged prescriptions). All non-robbery cases were eliminated from the study. At the experimental sites, 38 robberies were reported, while 56 robberies occurred at the control sites.[7] (See Table 1.)

Overall Clearance: When overall clearance rates (cleared by arrest plus exceptional) are compared, there is no significant difference ($x^2 = 1.62$, df = 1) between the experimental (68 percent) and control (55 percent) groups. However, the overall clearance rate (61 percent, or 57 of 94 cases) for the two groups represents an unusually high level of case solution. During the same period, July 1, 1976, through the end of April, 1977, only 37 percent of all reported armed robberies in Seattle were cleared (371 of 1,003 reports; source: SPD Monthly Crime Capsules). Part of the high clearance rate in the control group was due to the clearing of 18 cases (five by arrest; 13 by exception) through pictures taken at experimental sites; that is, pictures taken of robbers in experimental sites were identified by victims and witnesses in control site robberies.

If control-site robbery clearances which were caused by experimental-site pictures are deleted and clearance data reanalyzed, there is a statistically significant difference (see Table 2). While the experimental group retains its 68 percent clearance, only 34 percent of control cases were cleared without the aid of experimental site photographs.

---

[7] The higher robbery rate in control sites (three robberies for every four sites as compared to an experimental rate of two robberies for every four sites) is partially an artifact of when a site is designated as "experimental." Until a camera is actually in place, robberies that occur are not considered to be experimental robberies. Since initial placement and subsequent movement of cameras took approximately three months of the total 10.5 months, the total time at risk for the two groups is not equivalent.

Table 1. Robbery Case Clearance Rate by Site

	Experimental	Control	Total
Total robberies	38	56	94
Not cleared	12 (32%)	25 (45%)	37 (39%)
Cleared	26 (68%)	31 (55%)	57 (61%)
By arrest	21 (55%)	14* (25%)	35
Exceptional	5 (13%)	17** (30%)	22
Arrested for robbery at other experimental site	4	13	17
Arrested for robbery at site other than experimental/control	1	4	5

*Includes five cases in which suspects were identified and subsequently arrested through photographs taken at experimental sites. Exclusion of these cases results in nine, or 16 percent arrest rate.

**Includes 13 cases in which suspects were identified through experimental site pictures. Exclusion of these cases results in four, or 7 percent exceptional clearance rate.

Table 2. Revised* Robbery Clearance by Site

Clearance Status	Group	
	Experimental	Control
Cleared	26 (68%)	13* (34%)
Not cleared	12	25
Total	38	38*

*18 cases which were cleared because of experimental site photographs deleted.

Case Clearance by Arrest: When only cases cleared by arrest are examined, the difference between experimental and control group cases becomes more distinct. While 55 percent of all cases were cleared by the arrest of at least one suspect, only 25 percent of control site cases were cleared in the same fashion. This difference was highly significant ($x^2 = 8.87$, df = 1, $p < .01$). See Table 3 below.

Table 3. Robbery Cases Cleared by Arrest by Group

Case Cleared By	Group	
	Experimental	Control
Arrest	21 (55%)	14* (25%)
Other than arrest	17	42
Total	38	56*

*Includes five cases in which suspects were identified from pictures taken at experimental sites.

Robbery Suspects: While a total of 94 robberies occurred, the number of offenders involved was 126. Within the two study groups, 56 percent of experimental site robbers were arrested as compared to 22 percent of control site robbers (see Table 4 below). This difference was highly significant (x^2 = 15.52, df = 1, $p < .001$).

Table 4. Robbery Offenders by Group

Offenders	Group		Total
	Experimental	Control	
Arrested	27 (56%)	17 (22%)	44
Not arrested	21*	61**	82
Total	48	78	126

*Includes six identified suspects
**Includes 30 identified suspects

Reason for Arrest, Case Clearance: To determine the specific factor responsible for arrest and clearance data, the basis for each arrest was identified. See Table 5 below.

Table 5. Basis of Arrest by Group

Cause of Arrest and Clearance	Experimental		Control	
	Arrests	Clearance	Arrests	Clearance
Photograph	21 (78%)	15 (71%)	7 (41%)	5 (36%)
Arrest at or near scene	4 (15%)	4 (19%)	5 (29%)	4 (29%)
Victim/witness identification	1 (4%)	1 (5%)	2 (12%)	2 (14%)
All other	1	1	3 (18%)	3 (21%)
Total	27	21	17	14

Twenty of the 21 robberies cleared by arrest at hidden camera sites were photographed in progress. Fifteen (71 percent) of these clearances were due to photographs taken; four were the result of apprehension on or near the scene; one was due to identification by the victim; and one was due to other factors. Of the 14 robberies cleared by arrest at control sites, five were the result of photographs taken at hidden camera sites; four were the result of apprehension on or near the scene; two were due to identification by witness or victim; and three were the result of other factors.

In conclusion, robberies were significantly more likely to be cleared by arrest (55 percent versus 25 percent) in businesses in which hidden cameras were installed. Of the experimental site robberies, 71 percent of those cleared were due to the presence of photographs, as opposed to either other evidence or arrest on or near the scene.

Objective 2: To increase significantly the proportion of convictions for commercial robberies in which photographs are taken as compared with those commercial robberies not involving hidden cameras.

Data to evaluate this objective were obtained from the SPD Robbery Unit and the King County Superior and District Courts docket files.

Conviction Results: To determine if there is an increased conviction rate from the use of hidden cameras, a comparison was made between the number of arrests resulting in convictions for robberies committed within hidden camera sites and within control sites.

There were 27 arrests for robberies at hidden camera sites and 17 arrests at control sites. All arrests resulted in a determination of guilt except for six cases which were either pending or dispositions were unknown. Four of the six had outstanding warrants, and two arrestees were juveniles for whom court data were not available. See Table 6 below.

Table 6. Convictions by Group

	Arrests by Group	
Court Finding	Experimental	Control
Guilty	23	15
Other*	4	2
Total	27	17

*Comprised of two juvenile suspects in experimental group for whom data were not available; two adults arrested for both experimental and control robberies who "jumped bail."

Because of the unusually high conviction rate within the groups (100 percent conviction), it is impossible to say whether the presence of photographs would make a difference in conviction rates in more typical cases; e.g., in 1976, of 160 adult robbery defendants, 113 (71 percent) were found guilty of robbery, 12 (8 percent) were found guilty of lesser charges, and 35 (22 percent) were acquitted, dismissed or found not guilty (source: Seattle Police Statistical Report 1976, p. 45). However, of the 48 suspects in the experimental site robberies, the 23 convicted (48 percent) represent a significantly higher overall conviction rate than the 15 of 78 suspects (19 percent) in the control group ($x^2 = 11.61$, $df = 1$, $p < .001$).

Prosecutor Activities

While there were no differences between conviction rates for individuals for whom arrests occurred in experimental and control sites, the quality of the convictions may have differed. To examine this possibility, prosecutor actions were studied in terms of severity of recommended sentences and plea bargaining.

Prosecutor activities, rather than court-imposed sentences, were examined for several reasons. First, the prosecutor's goal was seen as being compatible with that of the police (i.e., conviction). Second, although not within the power of the project to affect the prosecutor's actions directly, it was assumed that the provision of robbery photographs would lead to more serious sentence recommendations and fewer instances of plea bargaining in order to obtain convictions because of the common goal of the project and the prosecutor.

Sentence Recommendation: Of the 38 total robbery cases with convictions, 74 percent of the convictions for the experimental group (17 out of 23) were obtained through pleas of guilty, compared to 80 percent for control group cases (12 out of 15). See Table 7 below. This difference in entering guilty pleas was non-significantly different ($x^2 = 0.19$, df = 1). In all instances of guilty pleas,[8] the prosecutor agreed to recommend less than the maximum possible sentence for all charges.

Table 7. Means of Obtaining Conviction

Guilty by	Group		Total
	Experimental	Control	
Trial	6 (26%)	3 (20%)	9
Plea of guilt	17 (74%)	12 (80%)	29
Total	23	15	38

To determine if the nature of sentence recommendations differed depending on whether photographs of the crime were available, they were classified into five types: (1) reduction in type of offense initially charged with; (2) recommending a sentence of shorter time than maximum possible; (3) both reducing initial charge and recommending shorter sentence (both 1 and 2 above); (4) dropping of either additional charges or counts and reducing recommended sentence length; and (5) reduction in type of offense initially charged with, reduction in recommended sentence length and dropping either additional charges or additional counts (1, 2 and 4 above).

When the type of recommendation was compared for the two groups of cases, there were no significant differences ($x^2 = 3.03$, df = 4, p = .55). (See Table 8.) If all recommendations involving reduction of initial charges are combined (types 1, 3 and 5 above), there remain no significant differences ($x^2 = 1.83$, df = 1, p = .18).

Analyses of conviction data do not show a significant difference between photographed and non-photographed robberies for either conviction rates (perhaps because of an abnormally high conviction rate within the control group) or sentence recommendations by the prosecutor in cases in which the offender pleaded guilty.

[8] Cases in which pleas of guilty occurred were used because they comprised the majority of cases, and the form "Statement of Defendant on Plea of Guilty" contained within the docket file includes the following statement: "I have been told that the prosecuting attorney will take the following action and make the following recommendation to the court: _____."

Table 8. Type of Sentence Recommendation by Group

Type of Plea Bargain	Group		Total
	Experimental	Control	
1) Initial charge changed	1 (6%)	1 (8%)	2
2) Sentence changed	4 (24%)	3 (25%)	7
3) Initial charge and sentence changed	5 (29%)	6 (50%)	11
4) Sentence changed and additional charges or counts dropped	6 (35%)	1 (8%)	7
5) Initial charge and sentence changed and additional charges or counts dropped	1 (6%)	1 (8%)	2
Total	17	12	29

Finally, only those cases in which the robber's weapon was visible[9] were examined to see if this may have influenced the decision to recommend less than the maximum possible sentence. Of 13 cases in which a weapon was clearly visible in the photographs, seven (54 percent) received less than the maximum possible sentence recommendation. When compared to control cases, this was a non-significant difference ($x^2 = 2.18$, df = 1, n.s.). However, given the small number of cases available for study, this non-significance may represent the unreliability of a small sample of true population difference.

---

[9] Robbery is defined as second degree unless the robber, either in the commission or immediate flight: (a) is armed with a deadly weapon, (b) displays what appears to be a deadly weapon or (c) inflicts bodily injury. If any of these occur, the offense is robbery in the first degree (RCW 9A.56.200 and 9A.56.210). First degree robbery is punishable by not less than 20 years imprisonment and/or not more than $10,000 fine, while second degree robbery is punishable by not more than 10 years and/or not more than $10,000 fine. Despite these lengthy sentences, the median length of stay for all robbers in Washington State adult institutions as of June, 1976, was 23 months. While judges are compelled to sentence convicted offenders to the maximum possible sentence (RCW 9.95.010), the Board of Prison Terms and Parole sets the minimum term of imprisonment and grants parole, which explains how the median length of stay for convicted robbers can be less than the maximum possible sentence. However, the Board is not allowed to set a minimum term of less than five years (RCW 9.95.040) if a deadly weapon was used in the commission of a robbery. Since the finding of guilt to commission of a crime while armed (RCW 9.41.025) restricts the Board's discretion, it was assumed that the prosecutor's office, as an advocate for the State, would attempt to remove dangerous offenders from society for as long a time as possible.

Plea Bargaining: To assess the project effect upon the use of plea bargaining, names of defendants in cases involving experimental and control group sites were submitted to the prosecutor's office. Based upon the King County Prosecutor's filing and disposition standards (a policy that determines sentence recommendations based upon prior felony convictions, type and nature of crime, multiple incidents and whether or not a weapon was used), a review of the cases was performed to determine if plea bargaining (defined as granting of concessions not granted by the sentencing standards) had occurred and whether it was a result of proof problems.

Of the 30 convicted offenders who were found guilty, five were involved in both experimental and control site robberies. Because plea bargaining is based on the defendant and not on the individual offense, these five persons (two of whom received plea bargains, both because of proof problems) were not included in the analysis of how the project affected plea bargaining.

Within the 16 offenders convicted of experimental site robberies, three (18.8 percent) received plea bargains, one (6.3 percent) of which was due to proof problems. Of the nine offenders convicted of control site robberies, three (33 percent) received plea bargains, two (22 percent) because of proof problems. These differences, while favoring project effect, did not reach statistical significance. However, such an interpretation (no project effect on plea bargaining) probably is due to the small number of cases upon which the present analysis is based.

Objective 3: To reduce significantly the cost of processing robbery cases from initial police response through investigation and prosecution and final court disposition for those cases involving hidden camera photographs as compared with other commercial robbery cases.

Project data, King County Superior and District Courts docket files, Seattle Police Statistical Report 1976 and the 1977 Police Department Budget were used as data sources in the following analyses. Two separate analyses were performed. The first examined the time spent in processing a case from arrest through conviction, and the second examined the cost to the Seattle Police Department budget to achieve a conviction.

The results of these analyses indicated the following:

1. Robbery photographs resulted in significantly reduced case processing time (0.95 months less).

2. It cost $1,228.41 to have a camera on-site to photograph a robbery (conservatively estimated so that errors would tend to overstate actual cost).

3. With photographs, investigative cost (detective-only) and victim loss totaled $811.74 to achieve a conviction of a robbery offender.

4. Without photographs, the investigative cost and victim loss in control sites totaled between $1,835.02 and $2,607.89 to achieve a conviction.

5. Experimental site convictions cost $2,040.15 (camera placement cost of $1,228.41 plus investigative and victim cost of $811.74). This cost is

between 22 percent less and 11 percent more than convictions in comparable robberies.

The procedures used to arrive at these conclusions are fully detailed below. Those interested in an initial overview may wish to go directly to discussion of Objective 4 and the analysis of project impact upon commercial robbery occurrence.

Case Processing Time: Arrest-to-conviction processing time was chosen for analysis because it was assumed that it should reflect the cost to the City in terms of both police response and investigatory efforts, and the cost of holding a suspect between the time of arrest and final disposition. As processing time decreases, there should be a corresponding decrease in police costs and in the cost of keeping suspects in jail. However, no estimates of potential cost savings were attempted because reliable data were judged to be unavailable. Processing time was determined for those cases in which the court outcome was known. Time was counted as the number of named months (e.g., January, February, March, etc.) from arrest to court disposition.

Twenty-three arrests at hidden camera sites had an average case processing time of 1.65 months, while the average processing time for the 15 arrests at control sites was 2.60 months (see Table 9 below). The difference in the amount of time elapsed in processing a case was significantly different between the two groups ($t = 2.45$, $df = 36$, $p = .02$)

Table 9. Processing Time Distribution in Months from Arrest to Conviction, by Groups

Number of Months between Arrest and Conviction	Number of Individuals by Group	
	Experimental	Control
0*	2	0
1	7	2
2	12	8
3	1	2
4	1	2
5	0	0
6	0	0
7	0	1

*Same month

This indicates that the presence of pictures of the crime being committed reduced the mean average processing time of cases resulting in conviction by 37 percent, or almost an entire month.

Cost of Investigation for an Arrest, Charge and Conviction: To examine actual processing cost, a comparison of experimental and control cases on the cost of making an arrest, obtaining a charge and achieving a conviction was performed.

There are many different ways to estimate personnel costs for an activity within the criminal justice system. Typically, costs are estimated on the basis of how

much time (and associated cost per unit of time) is spent performing the activity. However, this approach is accurate only if the total personnel time is productively spent (a situation that is rarely achieved in any work setting).

The approach used for this evaluation was to consider the robbery detectives as a resource whose sole purpose was the investigation of robbery cases. Using this approach, time engaged in any activity other than a "successful investigation" (defined as one resulting in a charge and conviction) is non-productive. This was felt to be appropriate because, if detectives did not perform this function, there would be no reasonable justification for their existence. Therefore, the cost/efficiency of the use of this resource will increase as either the number of successful investigations increases with the same resources, or the number of successful investigations remains the same with decreased resources.

Seattle's total 1976 robbery data are used as an example of the project cost-benefit analysis (see Table 10 below). The cost of the Robbery Unit within the Criminal Investigations Division (CID) was $361,744.[10] During 1976, 2,163[11] robberies were reported to the Seattle Police Department. Given the assumption that all cases were investigated and that the Robbery Unit exists only to investigate robberies, the department spent $167.24 on the investigation of each case (Robbery Unit budget/number of robbery reports, or 361,744/2,163). The mean average cost to each victim is conservatively estimated at $250.32, or the average value taken from all reported robberies.[12] This includes person robberies, which may be assumed to involve lower dollar loss than commercial robberies.

Table 10. Cost of Robbery Arrests, Charges and Convictions to Seattle Police Department Investigative Units and Victims; 1976

Item	Costs			
	Number of Reports Required to Produce One Item	Police Department Cost per Item*	Victim Loss**	Total SPD and Victim Cost per Item
Robbery report	1.00	$ 167.24	$ 250.32	$ 417.56
Adult arrest	7.05	1,178.32	1,764.76	2,943.08
Adult charge	11.27	1,884.08	2,821.11	4,705.19
Adult conviction	14.42	2,411.63	3,609.61	6,021.24

*Figured by dividing total Robbery Unit cost by total items

**Average loss of all robberies times the number of reports required to produce one item

---

[10] 1977 Annual Budget, City of Seattle, p. 534; cost based on (number of robbery unit/number of CID) detectives x CID total budget, or (12/95) x ($2,863,813).

[11] Seattle Police Department Crime Capsule: January through December, 1976, Seattle Police Department, dated January 11, 1977.

[12] Ibid.

Using the same sort of (total resource cost/number of activities) analyses, but using robbery arrests instead of robbery reports as the activity, during 1976, 307[13] adult[14] arrests occurred at a cost of $1,178.32 (Robbery Unit budget/number of adult arrests, or 361,744/307). On the average, 7.05 reports, involving victim loss of $1,764.76 (average loss times number of reports), occurred for each arrest.

In 1976, 192 adults were charged[15] at a cost of $1,884.08 per charge (total Robbery Unit budget/number of charges). For each charge of robbery entered by the prosecutor's office, there were 11.27 reports, with total victim loss of $2,821.11 reported. In 1976, 78 percent of known court dispositions for robbery[16] involved a finding or plea of guilt on the initial or lesser charges. The cost of the estimated 150 convictions (78 percent of 192) was $2,411.63 each to the department and $3,609.61 to victims. When both investigation costs and victim loss are added for each item, the cost for each robbery reported to police was $418; an adult arrest cost $2,943, an adult charge cost $4,705 and a conviction cost $6,021. It should be noted that the investigative costs are not additive. Each cost estimate for the activities (report, arrest, charge and conviction) includes within itself the cost for the other activities (e.g., the $167.24 report cost includes the cost of any subsequent arrest, charge and conviction cost to the Criminal Investigations Division).

Using the same procedure but restricting the analysis to experimental and control site robberies and using report, arrest, charge and conviction figures for these sites, the analysis was repeated.

Using 1976 police department cost for a robbery report (from Table 10, $167.24) and a different estimate of victim loss ($324.72[17]) as a starting point, relative police and victim costs were computed for control and experimental sites (see Table 11). Within the two groups of robberies which occurred in experimental and control sites, both the amount of victim loss and police cost generated by the number of cases investigated to produce an arrest, charge or conviction in experimental site robberies were substantially lower ($870.78,

---

[13] Seattle Police Statistical Report: 1976, "Adult Suspicion Bookings," Seattle Police Department, p. 49.

[14] Only adult robbery arrests, charges and convictions are dealt with because of the small number of juveniles involved and the fact that juvenile cases are handled by a different division of the Seattle Police Department.

[15] Seattle Police Department, loc. cit.

[16] Seattle Police Statistical Report: 1976, "Persons Charged 1976," p. 45. Only 160 case dispositions were available to the SPD statistical section. Of those known dispositions, 113 were guilty as charged, 12 guilty of lesser charges and 35 were acquitted or otherwise dismissed.

[17] Seattle Police Department, op. cit. Estimated victim loss was derived from armed robberies only (1,126, with a loss of $365,639) because it was felt to be more comparable with the commercial robberies under study.

Table 11. Costs of Arrests, Charges and Convictions to Police and Victims, by Group

Item	Group							
	Experimental				Control			
	Reports Needed per Item (a)	Police Cost* (b)	Victim Loss** (c)	Total Cost (d)	Reports Needed per Item (a)	Police Cost* (b)	Victim Loss** (c)	Total Cost (d)
Arrest	1.52	254.20	493.58	747.78	3.29	550.22	1,068.34	1,618.56
Charge	1.52	254.20	493.58	747.78	3.29	550.22	1,068.34	1,618.56
Conviction	1.65	275.95	535.79	811.74	3.73	623.81	1,211.22	1,835.02

*Based on 1976 figures for robbery reports ($167.24) times column (a)

**Based on average armed robbery loss in Seattle during 1976 ($324.72) times column (a)

$870.78 and $1,023.28, respectively--control total cost minus experimental total cost).

These figures indicate that much more productive use of investigation resources occurs when pictures of the robbery occurrence are available. However, the cost of obtaining those pictures must be included prior to making any final conclusions regarding cost effectiveness of the project.

Cost of Photographs: To determine the cost of obtaining the photographs in the experimental site robberies, project personnel costs, supplies and operating expenses, and initial equipment and eventual replacement costs were computed and then prorated for the time period for which data were available. All figures were computed conservatively so that all estimating errors should result in over-stating the cost of obtaining pictures of robberies-in-progress.

The procedure resulted in a maximum estimated cost of $1,228.41 per robbery. This was obtained by taking the annual project cost, $56,015.39 (see Table 12 for cost deviation) and multiplying this cost by 10/12, or the number of months the project was operational at the time of data collection. For this period of time, project prorated cost was $46,679.40. This cost was, in turn, divided by the number of robberies occurring within experimental sites (38), resulting in a cost of having a hidden camera on-site to photograph a robbery-in-progress of $1,228.41.

If one assumes that the most appropriate project objective is the conviction of offenders, the cost/benefit analysis of achieving convictions is $2,040.15 (cost of obtaining robbery photographs, plus the cost of investigation to achieve a conviction--from Table 11, Experimental Group, column [d]). Within a comparable group of stores (differing only on the basis of random assignment to either control--no camera or experimental--hidden camera status), the cost of achieving a conviction was $1,835.02 (from Table 11, Control Group, column [d]).

The cost difference for achieving a conviction was, at most, 11 percent higher in the hidden camera sites than in control sites. It should be remembered, however, that 23 of 48 (48 percent) robbery offenders within the 38 experimental

Table 12. Cost Estimates for Obtaining Photographs of Robberies

Item		Annual Cost
Personnel		
Detective*	$29,782.87	
Technician**	11,414.00	
Total Personnel cost, 12 months	$41,196.87	$41,196.87 (74%)
Supplies and Operating Expenses**		
(18 months)	$ 9,886.00	$ 6,590.67 (12%)
Equipment		
Initial purchase**	$28,700.00	
Replacement cost (estimated ten-year life; 7 percent compounded annual inflation)	56,457.24	
Subtotal ten-year cost	$85,157.24	
Salvage value of initial equipment: 10 percent	2,870.00	
Total ten-year Equipment cost	$82,287.24	$ 8,228.72 (14%)
		$56,015.39

*Estimated by dividing total 1977 CID budget by total number of detectives ($2,829,373/95 detectives). Project director's salary was paid by the Seattle Police Department.

**Taken from grant application.

site robberies were convicted while only 15 of 78 (19 percent) of robbery offenders within the control site robberies were convicted. In addition, an excluded factor in the cost analysis is that experimental site defendants required an average of a month less incarceration prior to conviction.

A further factor not taken into account in the above analysis is that five convicted offenders (involved in three cases) in the control group were initially identified through pictures taken at hidden camera sites. If these control cases were deleted from Table 11 and the police cost recomputed for 53 cases

(total control robberies [56] minus three cases in which five suspects were identified by project photographs) in which 10 convictions were obtained (15 total control convictions minus five in which suspects were identified through experimental-site photographs), the rate of the number of reported cases to achieve a conviction becomes 5.30, rather than 3.73. Using the same police investigation and victim loss figures as before ($167.24 and $324.72), the cost to achieve a conviction is $2,607.89. This cost figure would indicate that project conviction cost ($2,040.15) was 22 percent lower than comparable control conviction costs.

In summary, hidden camera site convictions are attained at from 11 percent more to 22 percent less police investigation and victim loss cost than convictions in comparable sites. These convictions are achieved in significantly less time (one month). In addition, arrests and subsequent convictions are much more likely to be attained (48 percent experimental versus 19 percent control offenders).

Objective 4: To reduce significantly the incidence of commercial robbery in the City of Seattle, as compared to other comparable jurisdictions.

This comparison was performed in two ways. The first comparison was made in the manner outlined in the project proposal. However, it was judged to be inappropriate because of the short time for which data were available following project implementation and the inclusion of data not relevant to the project. Therefore, a second analysis was performed which attempted to deal with these two problems.

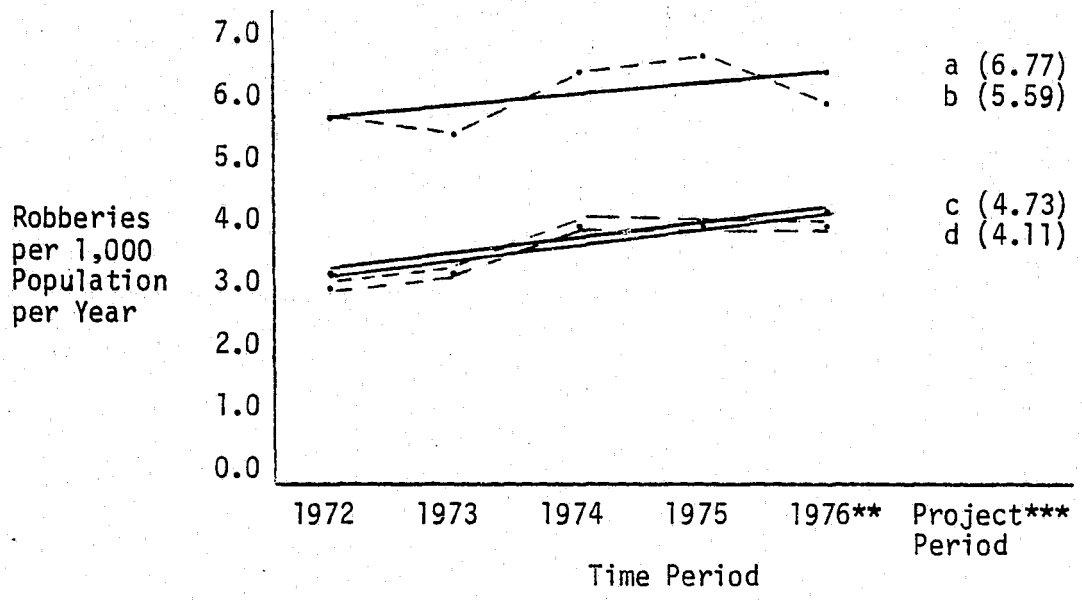
First Analysis: This objective was first assessed using data from the FBI's Uniform Crime Report System. Annual per capita robbery rates (all reported robberies per 1,000 population) were computed for all cities with 250,000 or more population. Similar rates were computed for Seattle using Seattle Police Department crime data and Seattle population estimates. Using 1972 through 1976 as base years, a linear regression prediction[18] was made for the project period July 1, 1976, through March 31, 1977, or the period for which data were available for both the United States and Seattle at the time of analysis. (See Figure 1.) While the 13.1 percent decline in Seattle's total reported per capita robberies was significant ($z = 4.68$, $p < .01$, based upon a test of actual versus predicted number of robberies as a proportion of the total Seattle population) from the first half of 1976 to the project period, comparable cities experienced a larger (17.4 percent) decline in robbery for the first three months of 1977 compared against the same three months in 1976.

---

[18] Linear regression prediction involves using the equation $y = a + bx$ to describe a relation between two variables. In this instance, y represents per capita robbery while a is a constant value, b is either a positive or negative weight to indicate the direction of trend as x changes, and x represents the year. Data used were as follows:

Year	1972	1973	1974	1975	1976	Project Period
U. S. rate	5.8	5.7	6.5	6.8	6.2	5.6
Seattle rate	3.0	3.3	4.0	4.2	4.2	4.1

Figure 1. Reported Robberies per 1,000 Population, Cities 250,000 and Larger* and Seattle



*Source of information: Computed from data in FBI reports (a) Crime in the U. S., table title, "Crime Rates, Offenses Known to Police, by Population Groups," (b) Uniform Crime Reports (1976 Preliminary Annual Release), and (c) Uniform Crime Reports (January through March, 1977).

**1976 annual rate computed on basis of January through June data for Seattle, for comparable cities on the basis of preliminary UCR release indicating a 9.78 percent decrease in cities over 250,000 population from 1975 to 1976.

***Project period for Seattle based upon July, 1976, through March, 1977, data; U. S. data based upon 9.16 percent decrease in cities of 250,000 population and larger for the period January through March, 1976, versus 1977.

This analysis cannot be viewed as being conclusive for several reasons. First, the data available for the project period cover only nine months (July, 1976, through March, 1977) and include non-commercial robberies which comprise about 75 percent of both United Stated and Seattle reports) which are not directly affected by project efforts. Second, while the project has been successful in identifying and convicting robbers, the majority of arrests were not obtained until late in the period available for examination. (See Table 13.) Given this, plus the average case processing time of 1.6 months, an attempt to find a

Table 13. Month of Arrest for Hidden Camera Cases Resulting in Conviction

Arrest Month	Number of Arrests	Arrest Month	Number of Arrests
June, 1976	1	November, 1976	5
July, 1976	0	December, 1976	2
August, 1976	0	January, 1977	2
September, 1976	1	February, 1977	5
October, 1976	4	March, 1977	3

reduction of crime occurrence through either an overall deterrence effect or reduction of robber population is probably premature. Third, strictly comparable time period data for the United States during the project period were not available at the time of analysis. National data for the project period were based upon a three-month percentage change for January through March, 1977 (using the same 1976 period as a base). Because of unavailability of data, the 1976 per capita data for the three-month period were estimated using preliminary percentage change data for the total 1975-1976 years. As a result of the double estimation of per capita data for comparable jurisdictions, the final estimates may be in error.

Second Analysis: Because of the various problems with the preceding analysis, the second analysis of the intent of Objective 4 (to reduce commercial robbery) was performed. First, to obtain more current data, only Seattle robbery data were examined, which provided four more months of information. Second, to control for historical trends or general changes in robbery rates, a non-equivalent control group design[19] was employed. In this analysis, non-commercial robbery data for Seattle were used as a comparison against commercial robbery data. This was based upon the assumption that while changes in the occurrence of commercial and non-commercial robberies are influenced by the same general factors (unemployment, social unrest, etc.), the offender populations for the two types of robberies are relatively distinct. Given this assumption, one would expect that a reduction in the number of commercial robbers would result in detectable reduction of commercial robberies while not influencing the number of non-commercial robberies.

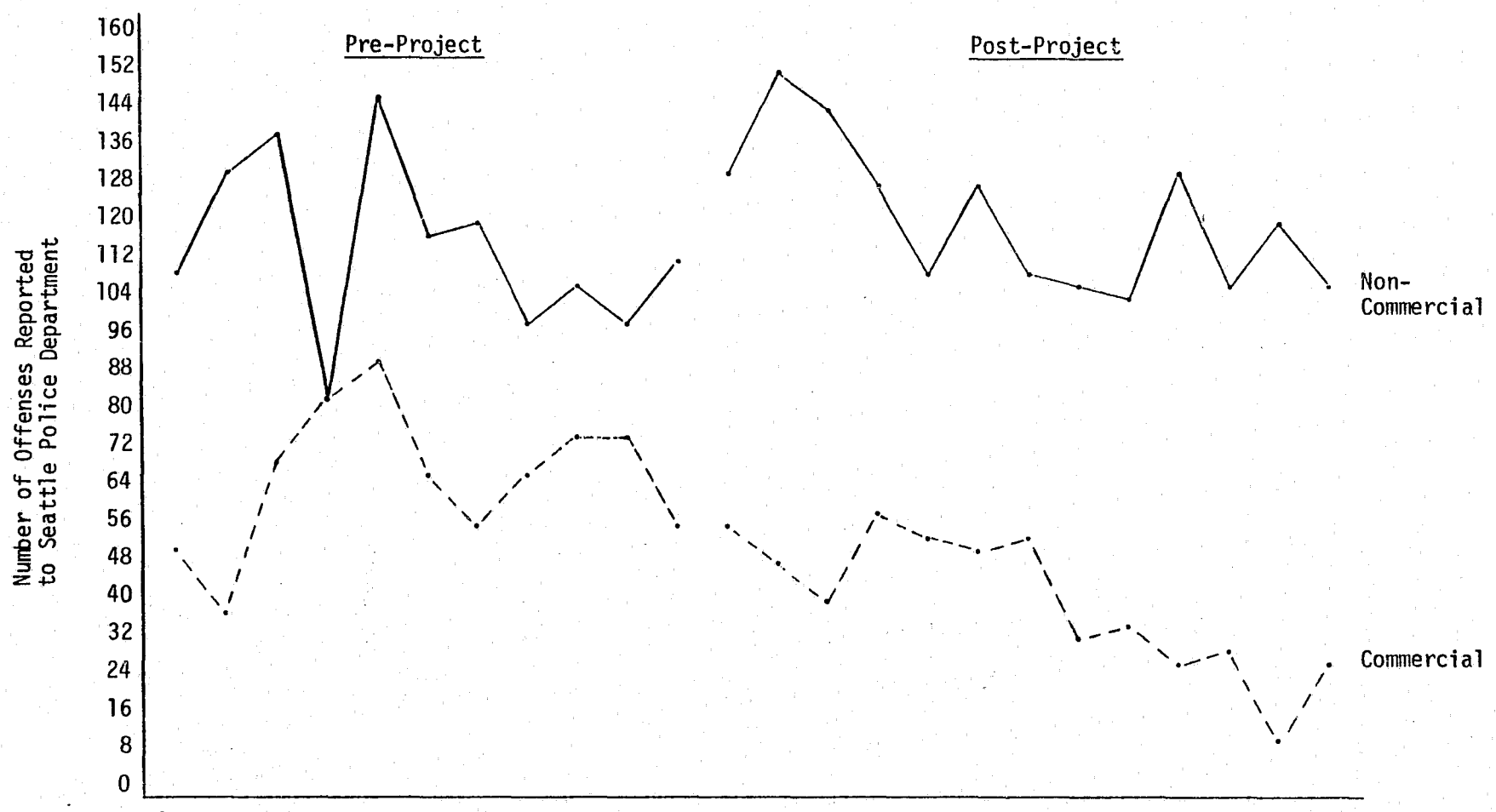
Data for this analysis (see Figure 2) were obtained from the SPD Data Processing Unit and covered the period August 1, 1975, through July 31, 1977.[20] Using August, 1976 (the first complete month following project implementation), through June, 1977 as a post-project period (July, 1976 and 1977 were not used because of a lack of corresponding "pre-" months), the corresponding months of

---

[19] D. T. Campbell and J. C. Stanley, Experimental and Quasi-Experimental Designs for Research, Chicago, Rand-McNally, 1963.

[20] Data prior to August, 1975, were not available because of absence of information needed to distinguish commercial from non-commercial robberies. (This required information for both the type of premises and sex of victim to distinguish a commercial robbery from a non-commercial robbery which occurred on a commercial premises; e.g., a tavern robbery as opposed to a robbery of an individual in a tavern or in a tavern parking lot.) July, 1977, was the last complete month for which data were available at the time of analysis.

Figure 2. Pre- and Post-Project Monthly Rates for Commercial and Non-Commercial Robberies



	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
	1975					1976												1977						
Commercial	49	34	70	81	93	66	60	67	73	73	58	58	46	37	63	57	53	56	32	33	27	30	9	27
Non-Commercial	110	130	139	81	146	116	119	100	108	100	112	132	152	146	129	109	129	111	107	103	133	107	119	107
Total	159	164	209	162	239	182	179	167	181	173	170	190	198	183	192	166	182	167	139	136	160	137	128	134

the preceding year (August, 1975, through June, 1976) were identified as the pre-project period.

To determine if the 38 percent reduction in monthly commercial robbery rate (from 65.8 per month between August, 1975, and June, 1976, to 40.6 per month between July, 1976, and July, 1977) was statistically significant, an analysis of covariance was performed. (See Table 14.) Same-named months in the pre- and post-project periods were paired as covariant and dependent variable data (e.g., August, 1975, with August, 1976), with commercial robbery and non-commercial robbery representing the two groups.

Table 14. Analysis of Covariance

Source of Variation	df	SSx	SP	SSy	Residuals		
					df	$MS\hat{y}$	F
Treatment	1	13,107.68	22,017.00	36,982.00	1	6,966.40	28.78
Error	20	5,948.18	2,229.64	5,434.36	19	242.03	
Total	21	19,055.86	24,246.64	42,416.36	20		

The results of the analysis ($F = 28.78$; $df = 1, 19$; $p < .001$) indicate that the decrease in commercial robbery rate (-38.8 percent) following program implementation is significantly different from the change in non-commercial robbery rates (+6.7 percent).

Additional support for the view that the project may have caused the decrease in commercial robbery can be obtained by correlating the cumulative number of persons arrested and convicted within camera sites (Table 13) with corresponding months' commercial robbery reports (Figure 2). There is a statistically significant negative correlation ($r = -.63$; $df = 9$; $p < .05$) which indicates that as the cumulative number of arrests has increased, the monthly rate of commercial robbery has decreased.

Further cause for attributing the change to project operation is the Robbery Unit's clearance of 248 robbery cases as a result of hidden camera evidence (robbery pictures leading to either arrest or additional charges, which in turn lead to offenders' admission of still additional robberies).

Objective 5: To accomplish project objectives without significantly increasing the risk of injury to victims, bystanders, police and offenders.

Data to evaluate this objective were obtained from the SPD Robbery Unit.

To determine whether project objectives were accomplished without increasing risk of injury, a comparison was made between the number of injuries occurring among victims, police officers and suspects/offenders within the cases examined. Of 93 cases examined, there were no injuries to either officers or offenders. However, there were four cases involving victim injuries (one injury for each case). Comparison of injuries at hidden camera and control sites indicates there were no significant differences in the number of injuries ($\chi^2 = 0.44$, $df = 1$, n.s.). (See Table 15.) Of those injuries that did occur, three were not serious enough to require any sort of medical attention. One case, a control-site robbery, involved the rape of a victim.

Table 15. Number of Injuries by Group

Injuries Occurred	Group: Hidden Camera Sites	Group: Control Sites	Total
Yes	1 (3%)	3 (5%)	4
No	37	52	89
Total	38	55*	93

*One robbery at a control site was eliminated because of unavailable data.

One unanticipated effect of robbery photographs did occur in one hidden camera site robbery. A suspect (a prior robbery convictee) was identified through "mug" shots by two victims as the offender in the case. However, the project photographs of the robbery proved that the initial suspect had not committed the robbery. Through the availability of the photographs, the mistakenly identified suspect was released and the actual offender was subsequently identified, arrested and convicted.

Additional Analyses: Offender Characteristics

A comparison of offenders' characteristics and past criminal histories was performed. This anlaysis was to answer the question of whether only naive, amateurish and generally inexperienced offenders would be foolish enough to get their pictures taken committing a robbery, or whether the covert operation of the project was good enough to capture both "professional" and "amateur" robbers. Seattle Police Department "rap," or local arrest history, sheets were used to examine whether convicted robbers differed on (a) whether they had ever been arrested before, (b) average number of arrests, and (c) severity of offenses for which they had been arrested. In addition, comparisons of offenders' age, race and sex were performed. For all of these comparisons, only those persons arrested, charged and convicted were used. Offenders were divided into three groups: those convicted of robbery in the experimental sites only, those convicted of robbery in control sites only and those convicted of robbery in both control and experimental sites.

Of the 25 individuals involved in the 27 arrests in experimental sites, two were juveniles (and excluded because of unavailability of data), seven were also convicted of robberies in control sites and two were each convicted of two separate experimental-site robberies. This resulted in 16 individuals convicted for 18 experimental-site robberies. Of the 16 individuals involved in the 17 arrests in control-site robberies, seven were also involved in experimental-site robberies, and one person was convicted for two separate control-site robberies. This resulted in nine persons convicted of ten control-site robberies. Of the seven persons arrested for robberies in both types of sites, two skipped bail, leaving five persons convicted of both experimental and control-site robberies.

Offender Characteristics Summary: When the three groups of convicted robbers are compared on the basis of past criminal history, there are no indications that would suggest that only amateur robbers have been arrested through project efforts. In fact, all analyses suggest that those photographed, subsequently arrested and convicted are likely to be more serious offenders than robbers who

are not photographed but who are arrested and convicted through more conventional means.

While the above conclusion may be challenged on methodological grounds of selection bias,[21] the practical results of project identification are not influenced. These results indicate that, first, many more robbers are identified, arrested and convicted through the use of hidden cameras; second, that those identified are more likely to be believed to have been involved or actually to have been involved in prior crimes.

The specific analyses performed to reach these conclusions are fully detailed below. Those seeking an initial overview may wish to go directly to the next section.

Age, Race and Sex Characteristics of Convicted Robbers: Within the three groups of convicted robbers, the age distribution was quite similar (see Table 16 below), with the exception being one elderly offender (age 73) in the experimental-only group. The difference in age of the three groups is non-significantly different ($F = 0.45$; $df = 2$).

Table 16. Age of Convicted Robbers by Site Robbed

Age	Group: Experimental (n = 16)	Group: Experimental and Control (n = 5)	Group: Control (n = 9)	Total
19	3	2	2	7
20	1	-	2	3
21	2	-	1	3
22	1	-	-	1
23	2	-	-	2
25	2	1	1	4
27	2	1	1	4
28	-	-	1	1
32	-	-	1	1
33	1	1	-	2
43	1	-	-	1
73	1	-	-	1
Median age	23	25	21	23
Mean age	27.50	24.60	23.64	25.80

When race of robbery offenders was examined (see Table 17), it was found to differ significantly by group ($x^2 = 8.90$, $df = 2$, $p < .02$). However, when experimental-only are compared with control-only offenders, the difference is

[21] Specifically, robbers photographed but not arrested during the initial patrol officer response are increasingly likely to be identified by detectives the more times the robber has been arrested in the past. The effect of this would be to produce a group of identified robbers who were more "serious" offenders than the total group of photographed robbers, while the non-identified but photographed robbers would include less "serious" offenders and more persons who had never been arrested locally.

Table 17. Race of Convicted Robbers by Site Robbed

Race	Group			Total
	Experimental	Experimental and Control	Control	
White	13 (81%)	1 (20%)	8 (89%)	22 (73%)
Black	3	4	1	8

non-significant (x^2 = 0.25, df = 1, n.s.), indicating that the initial difference is due to the presence of the offenders arrested and convicted of offenses in both types of sites.

All convicted offenders were male.

Any Past Arrests: An arrest was counted for each separate physical booking or citation that appeared within the individual's local rap sheet. Rebookings on old charges were not counted unless they included new offenses. As such, this analysis counts the number of times the individuals were investigated, but not the number of reasons for which they were investigated.

Of those convicted of robberies in only experimental sites, 69 percent had prior arrests and/or citations listed on local rap sheets (see Table 18 below), while 44 percent of those convicted in control-only sites had prior arrests. Of those convicted of robbery in both types of sites, 80 percent had prior arrests. However, these differences were not statistically significant (x^2 = 2.16, df = 2, n.s.). Given the small sample size, the lack of significance could represent either a lack of real difference or unreliability of the sample because of the few cases.

Table 18. Arrest History by Group

Frequency of Prior Arrests and/or Citations	Group			Total
	Experimental	Experimental and Control	Control	
0	5 (31%)	1 (20%)	5 (56%)	11 (37%)
1	0	0	1	1
2	3	0	1	4
3	1	0	0	1
4	0	1	2	3
5	2	0	0	2
6	2	1	0	3
8	0	1	0	1
10	1	0	0	1
13	1	0	0	1
16	0	1	0	1
19	1	0	0	1
Total arrests/citations	73	34	11	118
Mean average arrests/ citations	4.56	6.80	1.22	3.93

Average Number of Prior Physical Arrests and Citations: While the average number of prior arrests and/or citations was quite varied for the three groups (4.56 per offender in the experimental-only, 6.80 in the experimental-and-control and 1.22 in the control-only group), the differences were not statistically significant by one-way analysis of variance (F = 2.54, df = 2, 27).

Again, this may represent either a true lack of population differences or unreliably small samples.

Offense Severity of Past Reasons for Arrest: In the preceding section, the number of separate instances of arrest and citations were examined. In the following section, the separate number of charges or reasons for arrest or citation are analyzed. Since individuals were arrested for multiple reasons at one time, the following data will indicate a higher rate of offenses than the preceding section.

For the present analysis, offenses were defined as serious if they were FBI Part I crimes (see Table 19 below). Within the experimental group, there were 19 prior Part I arrest reasons (1.19 per offender), while the control group had eight prior Part I arrest charges (0.89 per offender). However, the highest rate was achieved by those arrested and convicted for robbery in both experimental and control sites, 22 (or an average of 4.40 prior Part I arrest charges per offender). When total reasons for arrests and/or citations are examined, the same ranking is found. Those convicted of robberies in both types of sites had the highest average (12.40, of which 39 percent were traffic or traffic-related), followed by experimental-site robbers (9.69, of which 52 percent were traffic), then control-site robbers (2.38, of which 11 percent were traffic).

Table 19. Past Reasons for Arrest by Group

Offense	Group			Total
	Experimental (n = 16)	Experimental and Control (n = 5)	Control (n = 9)	
Part I Offenses*	19	22	8	49
Negligent manslaughter	2	0	0	2
Robbery	8	11	1	20
Assault	1	4	0	5
Burglary	0	1	2	3
Larceny	4	6	5	15
Auto theft	4	0	0	4
Other	55	16	9	80
Traffic	81	24	2	107
Grand Total	155	62	19	236
Number of persons with one or more Part I arrest	8 (50%)	4 (80%)	3 (33%)	15 (50%)
Number of persons with prior robbery arrest	6 (38%)	4 (80%)	1 (11%)	11 (37%)

*No prior arrests for homicide or rape were present in the groups' history.

When these data are analyzed on the basis of whether or not individuals had been arrested for one or more prior Part I charges, there were no significant differences ($x^2 = 2.80$, df = 1, n.s.).

If these data are analyzed using statistically more powerful tests (ANOVA), there are significant differences in the mean number of prior Part I arrest charges ($F = 6.36$; $df = 2, 27$; $p < .05$). However, these differences are not due to statistically reliable differences between the experimental-only and control-only robbers ($t = 0.43$, $df = 23$).

When analyzed on the basis of whether they had been arrested in the past in connection with robbery investigations, there were significant differences among the three groups ($x^2 = 6.58$, $df = 2$, $p < .05$). While not statistically reliable because of the small sample size, a comparison of number of prior robbery arrestees between experimental-only and control-only robbers indicates no difference ($x^2 = 1.99$, $df = 1$).

Summary

Evaluation of the Hidden Cameras Project after almost a year (10.5 months) of field operation indicates the following:

1. The project was able to photograph 84 percent of the 38 robberies occurring in the target sites. Each of the 75 cameras can be expected to be in position to record a robbery every 1.7 years.

2. The clearance rate in experimental-site robberies (68 percent) was significantly higher than that of control-site robberies (34 percent) caused by conventional means. An additional 21 percent of control-site robberies (for a total of 55 percent) were cleared by arrests or identifications brought about through photographs taken at experimental-site robberies.

3. Arrest data show clear and statistically significant differences. While 55 percent of all experimental cases were cleared by arrest (20 of the 21 cleared cases being photographed), only 25 percent of control cases were cleared by arrest. Of the 48 offenders in experimental cases, 56 percent were arrested, while of the 78 control site offenders, only 22 percent were arrested.

4. Conviction rates were not examined as a function of whether photographs of the crime were available because of an abnormally high rate for both groups (100 percent conviction for all adult cases disposed of). However, significantly more of the robbers in the experimental group (48 percent) were eventually identified, arrested and convicted than was true in the control group (19 percent).

   Entering a plea of guilty was as likely to occur in experimental cases (74 percent) as in control cases (80 percent). Analysis of severity of prosecutor recommended sentences and frequency of engaging in plea bargaining (18.8 percent for experimental and 33.3 percent for control) were non-significantly different.

5. The presence of robbery photographs resulted in cases being processed from arrest to conviction significantly faster (1.65 months, versus 2.60 months).

6. Using conservative cost estimates (tending to overstate project costs) resulted in an estimate of $1,228.41 to have a camera present during a robbery. Using average victim loss and average detective case investigation costs only (which excludes the average one-month savings in jail costs between arrest and conviction), project-conviction cost was estimated to be $811.74. Including photographic costs, total project-conviction cost was $2,040.15. Estimated conviction cost in directly comparable cases without hidden camera photographs was between $1,835.02 and $2,607.09. This represents a range (conservatively estimated) in which project-conviction costs were from 22 percent lower to 11 percent higher than conventional investigation costs.

7. The decline in all reported robberies (both commercial and non-commercial) following project onset was not significantly lower than that of comparable cities. However, given various methodological problems (see full text), this was not judged to be a conclusive indication of project effect.

   A more precise analysis of just commercial robbery data using local non-commercial robbery rates as comparison data resulted in finding a statistically significant 38.8 percent decline in commercial robbery following project onset, while non-commercial robberies increased by 6.7 percent. The decline in commercial robbery was found to be significantly correlated with the number of robbers arrested and convicted during the project period ($r = -.63$, $p < .05$).

8. Project objectives were achieved without significantly increasing risk to either victims, police or offenders. In fact, the presence of photographs prevented one ex-offender from being wrongfully charged on the basis of eyewitness testimony.

9. Persons arrested and convicted as a result of project photographs are not less "serious" offenders in terms of past criminal history than those arrested through other means. There are indications that the project may, in fact, identify more serious offenders as indicated by local arrest history.

SECTION 4.B.

BELLEVUE CITIZEN INVOLVEMENT IN BURGLARY PREVENTION GRANT EVALUATION*

By

City of Bellevue
King County
Law & Justice Planning Office

Tony Mulberg

and

Shelley Wein

*Selections from the Bellevue Citizen Involvement in Burglary Prevention Evaluation have been excerpted for inclusion.

INTRODUCTORY COMMENTS ON:

"BELLEVUE CITIZEN INVOLVEMENT IN BURGLARY PREVENTION"*

The following is taken from an evaluation by Tony Mulberg and Shelley Wein, "Bellevue Citizen Involvement in Burglary Prevention." This study is interesting on several counts: It confronts difficult but typical problems, and the authors have carefully attempted to control these problems; yet several problems remain which illustrate possible avenues for further analysis. The major problem these evaluators faced was that the burglary prevention program was undertaken for the entire city precluding the use of a treatment versus control group comparison in the research design. To add to the problems of evaluation, the city also instituted a team policing policy at approximately the same time, making it impossible to be sure any observed effects should be attributed entirely to the burglary prevention program. Mulberg and Wein's approach was to use a regression discontinuity analysis for monthly time series of residential burglary. Although this design has several weaknesses noted by the authors, this analysis provided an overall assessment of the program on a city-wide basis. Because two neighborhoods in Bellevue were targeted for extensive application of the burglary prevention program, Mulberg and Wein were able to conduct a second analysis in which they compared the effects of extensive treatment to the effects of less intense application of the program in the rest of the city. This evaluation report is followed by a rather extensive discussion that demonstrates how different types of time series analyses strengthen and illuminate the conclusions drawn by the author.

*This review was prepared by Dr. Jerry Medler.

BELLEVUE CITIZEN INVOLVEMENT IN BURGLARY PREVENTION

GRANT EVALUATION

Tony Mulberg & Shelley Wein

I. Introduction

A. Background.

The concept of citizen involvement in burglary prevention has received increased attention from the law enforcement community over the past few years. In response to the increasing frequency of reported residential burglary, and the apparent amenability of burglary to prevention techniques, numerous burglary prevention programs involving citizens have been organized and implemented.

These burglary prevention programs have emphasized neighborhood meetings at which law enforcement or civilian crime prevention personnel discuss: (1) the specific neighborhood's burglary problem, (2) the concept of neighborhood block-watch, (3) the importance of marking personal property for positive identification, and (4) the use of proper security devices within the home.

The City of Bellevue, having realized a need for a burglary prevention program, applied for Law Enforcement Assistance Administration (LEAA) funds to implement a prevention program. The project was approved by the Governor's Committee on Law and Justice, and the State Office of Community Development (i.e., Law and Justice Planning Office) awarded the contract. The first year grant covered the time periods of May, 1974, through April, 1975.

Bellevue implemented the program in May, 1974, in two matched areas of the city. One area was designated as a target area, while the other area was designated a comparison or control area. The target area received extensive burglary prevention services, including police sponsored neigh-

borhood meetings on crime prevention, block-watch organization, access to engravers for marking personal property, and home security checks by the police. The comparison area received virtually none of the treatments.

Near the end of the project's first year of oepration, a second grant application requesting funds to expand the burglary prevention program city-wide was submitted and approved. Beginning May 1, 1975, a greater emphasis was placed on involving as many citizens in the program as possible. This was to be accomplished through a mass media campaign and lectures at civic meetings. Door-to-door campaigning was paramount to the program's success and was instituted initially in the high crime areas of Bellevue (identified as Ardmore and Enatai).

The second year of the burglary prevention project grant concluded April 31, 1976. This evaluation report reflects the finds of the second year evaluation (May 1, 1975, through April 31, 1976) conducted by the King County Law and Justice Planning Office. Third year continuation funding was not requested by the City of Bellevue. The Bellevue Police Department, however, did institute a team policing project which continued the prevention program as part of the team policing strategy.

The goals of the Bellevue burglary prevention program were to: (1) bring to the attention of the public the seriousness of the city's burglary problem, (2) inform the public of burglary prevention measures, and (3) encourage the public to institute the recommended burglary prevention safeguards. The ultimate goal of the project was, of course, a reduction in reported residential burglary.

D. Limitations of the scope of this report:

There are several limitations on the scope of this study which must be

considered when reviewing the findings.

1. Weak research design. A before/after design and a non-equivalent control group design were used in this study. As with all quasi-experimental research designs, internal and external validity factors undermine the ability to directly attribute any decline in reported residential burglary to the program activities alone. Some of the validity problems with the pre-post and non-equivalent control group designs include self-selection bias, multiple treatment effects, and statistical regression. These questions of validity must be considered when interpreting findings regarding program success.

2. Reported versus unreported crime. A continual problem with crime statistics is that not all crimes are reported to the police. As has been documented in numerous victimization studies, large numbers of crimes go unreported to the police.[1] A victimization study completed by the City of Seattle found a burglary reporting rate of only 46%. This means that almost half of the burglaries in Seattle went unreported to the police. Accordingly, crime statistics used by evaluators may be incomplete. One elment further complicating the validity of reported crime rates is that those participating in program treatments may be more inclined, as a result of the program, to report crimes to the police. An increase in the number of reported residential burglaries after program implementation may reflect an increased tendency to report burglaries to the police, not an increase in the actual number of burglaries (or a combination of the two). Unless a victimization survey is conducted, it is impossible to determine whether

---

[1] Schram, D.C. Study of Public Opinion and Criminal Victimization in Seattle, City of Seattle Law and Justice Planning Office, 1973. Schneider, Anne L. Evaluation of the Portland Neighborhood Based Anti-Burglary Program, Oregon Research Institute, March 1975.

an increase in the number of burglaries is attributable to increased reporting or an increase in actual burglaries.

Seattle's Law and Justice Planning Office with grant funds from the State Law and Justice Planning Office conducted Bellevue's first-year "Citizen Involvement in Burglary Prevention" victimization study of Bellevue's crime reporting rates. Unfortunately, errors in coding of survey data occurred and actual crime rates still remain unknown. (It is anticipated that the results of the Bellevue victimization study will become available within the next few months.)

3. Insufficient follow-up time. In this study, the time periods of one year after the program starting date were used. This allowed for only a one-year follow-up of the "treatment" area. Also, not all treatment areas received treatment during the first months of the project, i.e., some treatment areas did not receive treatment until late in the project year. Actual program effects on participants with minimal follow-up time may not become apparent until several months from now.

For example, one large area of Bellevue (Woodridge) did receive an intensive door-to-door campaign, but it was not begun until December 1975. While numerous citizens in the Woodridge area had been exposed to the prevention measure, including block-watches, property marking, and home security inspections by April 31, 1976, not enough follow-up time had passed to include this area in the treatment group.

4. The number of prevention program participants were slightly underestimated. The number of burglary prevention program participants were slightly underestimated. For example, not all citizens who attended a neighborhood meeting signed the attendance sheet. In addition, citizens

who did not attend a neighborhood burglary prevention meeting could still participate in a block-watch after receiving a block-watch briefing from a participating neighbor. As a third example, some project records were not maintained consistently during the project period. As a result, the actual number of prevention program participants could not be fully documented.

5. Small numbers. While the sample areas of Ardmore and Enatai represented 6.2% of the Bellevue population and reported 8.1% of the residential burglaries in the pre-period, the actual number of reported residential burglaries in these two areas is very small. Besides the small numbers leaving little room for improvement (i.e., there is little room for improvement at ten burglaries per month) or being subject to minor changes in the activities of burglars, small numbers are easily influenced when converted to percentages or subjected to statistical analyses. The numbers involved in these two areas then may be too small to allow firm conclusions about the effect of the program.

II. Methodology

A. Research design.

As noted in section I-D, two quasi-experimental designs were used in this evaluation. The pre/post test design was used to evaluate the impact of the prevention program on a city-wide basis, while the non-equivalent control group design was used to evaluate the effectiveness of the program elements for participant neighborhoods versus non-participant neighborhoods.[2]

[2]The terms 'participant' and 'non-participant' were used relatively in this report. Participant areas were defined as those areas in which: (1) an

The reader should again note the internal and external validity problems with these two evaluation designs (see section I-D).

B. Data collection and analysis.

Primary data sources for the evaluation consisted of Bellevue Police Department records, and burglary project records.

. . .

E. Impact of the Bellevue Burglary Prevention Program on reported residential burglary.

For the purpose of this evaluation, the restated objective against which the program was evaluated was:

> Given the operation of the residential burglary prevention program in Bellevue, a statistically significant decrease will be documented when the number of reported residential burglaries before the program is compared to the number of reported residential burglaries after program initiation.

To determine the program's success in achieving its objective to significantly reduce the incidence of residential burglary two measures were used on a city-wide basis, and two measures were used for a participant neighborhood, and non-participant neighborhood comparison.

The two measures of achievement used on a city-wide scale were:

(1) The number of reported residential burglaries for pre- and post-

---

intensive door-to-door campaign was done by the prevention program staff, and (2) a large number of households in the respective area participated in one or more of the prevention program elements. Conversely, a non-participant area was defined as an area in which: (1) minimal or no door-to-door campaigning was done by the prevention program staff, and (2) minimal or no households in the respective area participated in the prevention program elements.

program months were compared by means of a regression discontinuity analysis, and

(2) The rates of reported residential burglary (per 100 households) for pre- and post-program initiation months were compared by means of a t-test.

The two measures of achievement used for the participant neighborhoods' and non-participant neighborhoods' comparison were:

(1) The monthly rates of reported residential burglaries (per 100 households) were compared for participant and non-participant areas for pre- and post-period months by means of a t-test, and

(2) The number of reported residential burglaries for pre- and post-months were compared by means of a regression discontinuity analysis performed independently for participant and non-participant areas.

1. City-wide Analysis. As explained earlier, the pre-program months used were May, 1974, through April, 1975, or the year prior to the prevention program's implementation city-wide. The post program months used were May, 1975, through April, 1976, or the year following the city-wide program implementation date. Bellevue did have the first year burglary prevention grant project operating during the year May, 1974, through April, 1975. Some citizens in the target neighborhood were exposed to the burglary prevention strategies. However, the experimental area used in the first year grant project experienced only 28 reported residential burglaries during the year May, 1974, through April, 1975. Even if the program had been 100% effective it would not have affected the city-wide reported residential burglary frequency. Therefore, second year project was minimal at best, the year May, 1974, through April, 1975, was sufficient as a baseline data

year.

a. Measure one - regression discontinuity analysis. Graph 2 indicates the results of the city-wide regression discontinuity analysis. According to Campbell,[2] this test is appropriate when services cannot be denied to a control group, as was the case in Bellevue. The methodology is as follows: the least squares regression equation is computed on the basis of the number of burglaries in the pre-program months; the least squares regression equation is also computed for the post-program months. The regression lines are then plotted and compared; substantial differences are demonstrated when the slope and the intercept of the two regression lines differ.

Graph 2 indicates that the number of reported residential burglaries were substantially reduced during the post-program initiation period. The equations compared as follows:

	Slope		Intercept
Pre-program Y =	2.688(x)	+	56.11
Post-program Y =	0.033 (x)	+	62.88

The regression lines were different in both intercept and slope. While the slope in the post-program initation period was not negative, (which would have indicated a decreasing trend) the number of reported residential burglaries was down and appeared to have stabilized.

Since reported residential burglary figures were used for the purposes of the evaluation, a bias may have been introduced, as reporting rates may

---

[2] Campbell, Donald. "Reforms as Experiments," American Psychologist, Vol. 24, No. 4, (April, 1969), pp. 409-429.

tend to increase during prevention program operation.[3] As a result, the actual numbers of reported residential burglaries may be over-represented in the post-program initiation period.

b. Measure two - t-test. During the pre-program period the mean monthly rate of reported residential burglary (per 100 households) was 3.64. During the post-program initiation period the mean monthly rate of reported residential burglary (per 100 households) was 2.91. A t-test for significance comparing the monthly rates in the pre-program period to the post-program initiation period indicated that the decrease during the post period was statistically significant ($p<.025$, see Appendix VI, table 4).

Limits to these types of analyses however, preclude attributing the decrease in the number and rate of reported residential burglaries to the program elements alone, i.e., there may be other factors unrelated to the program contributing to the decrease. Other possible explanations for the decrease include:

(1) Statistical regression towards the mean. The regression effect (Campbell, 1969) suggests that Bellevue's reported burglary level was uncharacteristically high during the pre-period and therefore "artificial". The further suggestion by Campbell is that an "artificially" high level (as was the case for burglary in Bellevue) would abate regardless of any intervention strategies (e.g., Bellevue's Burglary Prevention Program). This explanation has some validity as the pre-project year of May, 1974, through April, 1975, did realize a greater incidence of reported residential burglary in Bellevue than any other year over the last five years.

---

[3] Schneider, A. L. Evaluation of Portland Neighborhood-based Anti-Burglary Program. Oregon Research Institute, 1975 (pp. 7, 16).

(2) Bellevue Team Policing. In June, 1975, the Bellevue Police Department implemented a team policing strategy. Under this strategy, patrol officers were given the added responsibility of performing functions which in the past were normally performed by detectives (e.g., follow-up investigations, gathering of evidence, etc.). For those cases followed-up, increases in both the arrest and clearance rates were anticipated. With the police officers aware of the burglary prevention program in existence, special emphasis was given the burglary case follow-ups.

The data in Appendix VII, Table 5, show that burglary "arrest rates" have increased from the pre- to post-prevention program periods.[4] Rates of burglary arrests per month over the number of burglaries per month were computed for the pre- and post-burglary project periods. The arrest rate increased from 18.91% in the pre-period to 24.66% in the post-prevention program period. A t-test performed on the monthly rates for the pre- and post-periods indicated that the increasing arrest rates for the post period were not statistically significant ($p.<05$) although, the changes in arrest rates may be considered significant in a practical sense.[5]

The implication is that team policing strategies may have impacted the prevention program, i.e., contributed to the reported residential burglary decrease, as burglars have either (1) been removed from the street by arrest, jail and/or prison, and are therefore unable to commit burglaries, or

---

[4]The "arrest rate" is not a true arrest rate in that the arrests during any given month are not necessarily those for burglaries committed during the same month.

[5]The calculated value of t approached significance at the .05 level (t=1.667, 22df.

(2) burglars became aware of the new police strategies and were, therefore, deterred from burglary activities. Most of the newspaper articles about the burglary prevention program also made reference to the team policing strategy; the burglary statistics then may reflect the combined effects of the two programs.

Initially, four officers were assigned to do burglary prevention presentations, while two student interns and a citizen volunteer performed the necessary door-to-door compaigning. By the end of December, 1975, 41 officers had been trained to give the presentations while 17 officers (excluding the staff) were actually involved in giving the program presentations. When the student interns completed their internships in August, 1975, and the citizen volunteer departed in November, 1975, the Bellevue police officers took over the door-to-door campaign; contacting citizens and setting up neighborhood meetings. The implication here is that many more neighborhood meetings were held and therefore block-watches formed as a result of the increased use of team police officers functioning as burglary prevention program personnel. Team police officers were responsible for 23 of the 89 block-watches organized, or 25.84% of the neighborhood block-watches.

2. Comparison of participant and non-participant neighborhoods. While it was clear that Bellevue did experience a statistically significant decrease in the number and rate of reported residential burglary after the implementation of the burglary prevention program, it was not possible with the measurements used, to fully attribute the decrease to the operation of the burglary prevention program. By a comparison of intensive participant areas with minimal participant areas in the amount of burglary reduction after the

prevention programs implementation, some inferences could be made regarding the possible effectiveness of the burglary prevention program.

As noted earlier, sections of Ardmore and Enatai (referred to as Area One) received extensive door-to-door campaigning and the program participation response by citizens in these areas was very favorable.* Within the participating sections of Ardmore and Enatai 566 (35.64%) of the households in this area participated in a block-watch, and 400 (25.14%) of the households in this area participated in the property engraving/decal element of the program. These two high participant areas were combined and compared to the rest of Bellevue, (referred to as Area Two excluding the sections of Ardmore and Enatai) which had received only minimal exposure to the prevention program, and in which the participation response was minimal. In the Bellevue area, (excluding Ardmore and Enatai) only 802 (3.29%) of the households in this area participated in a block-watch, and only 129 (0.53%) of the households in this area participated in the property engraving/decal program element.

During the pre-program period, Area One reported 72 residential burglaries in Bellevue. During the post-program initiation period, Area One reported only 37 burglaries, representing 4.89% of the total reported residential burglaries. During the pre-program period Area Two reported 811 residential burglaries representing 91.85% of the total reported residential burglaries in Bellevue. During the post-program initiation period, Area Two reported 720 residential burglaries representing 95.11% of the total

---

*The section of Ardmore referred to in this report was bounded by the following streets: North - 24th St. N.E., South - Northrup Way, East - 170th Ave. N.E., West - 156th N.E. The section of Enatai referred to in this report was bounded by the following streets: North - Decar Crest Lane and Parkridge Lane, South - 26th, East - 108th S.E., ...

reported residential burglaries in Bellevue. This reflects a 48.61% (from 72 to 37) decrease in the number of reported residential burglaries for Area One in the post-prevention program initiation period as opposed to an 11.22% (from 811 to 720) decrease in reported residential burglary experienced by Area Two. (See Table 6.)

TABLE 6

Frequency and Percent of Bellevue's Total Reported Residential Burglaries By Area

	PRE		POST	
	F	%	F	%
Area One	72	8.15	37	4.89
Area Two	811	91.85	720	95.11
	833	100.00	757	100.00

a. Measure one - t-test for significance

The mean monthly rates of reported residential burglary (per 100 households) further supported the apparent decreases in reported residential burglary frequencies. The mean monthly rate for Area One in the pre period was 4.81. The mean monthly rate in the post-period was 2.33. A t-test for significance indicated this rate decrease in Area One was statistically significant ($p<.005$, see Appendix IX, Table 7). The mean monthly rate of reported residential burglary (per 100 households) for Area Two dropped from 3.57 in the pre-period to 2.95 in the post-period. However, a t-test for significance indicated this rate decrease was not statistically significant ($p>.05$, see Appendix IX, Table 8).

b. Measure two - regression discontinuity analysis

One more attempt was made to measure the statistical significance of the changes in reported residential burglaries for the intensive citizen participant areas and the minimal citizen participant areas. The regression discontinuity analysis was used for each area independently.

The least squares regression line for Area One in the pre-program period (May 1974 through April 1975) indicated an increasing slope, the numbers of reported residential burglaries were increasing. During the post program initiation period the computed least-squares regression line indicated a decreasing slope, hence the numbers of reported residential burglaries were decreasing. The projected number of reported

residential burglaries for May 1976 for Area One was zero. As can be seen, the slope and the intercepts are different (see Appendix X, Graph 3).

		Slope		Intercept
	Pre - Y =	.280(x)	+	4.182
Area One	Post - Y =	-.507(x)	+	6.377

Area Two in the pre-period indicated an increasing slope, hence reported residential burglaries were increasing. During the post-program initiation period the computed least squares regression line while different in slope and intercept from the pre-period, indicated that reported residential burglary was still on the increase, although the actual frequency of reported residential burglary was down. (see Appendix X, Graph 3.)

		Slope		Intercept
	Pre - Y =	1.989(x)	+	54.652
Area Two	Post - Y =	.371(x)	+	57.591

In summary, the only decreasing trend was experienced in Area One during the post program initiation period. This suggests that Area One (Ardmore and Enatai) which received the greatest amount of burglary prevention services, and maintained a high citizen participation rate, experienced the greatest decrease in reported residential burglary frequency and rate.

It is not known to what extent, if any, displacement effects may have influenced the project outcome. The usual assumption about displacement is that it is most apt to occur in areas close to the experimental area, in this case Area One. As no specific residential burglary comparisons were made of those neighborhoods surrounding the Ardmore and Enatai areas, no inferences were mde regarding the transference of residential burglary activities after the burglary prevention program was implemented.

IV. Summary and conclusions

This evaluation report has analyzed:

(1) descriptive data covering the characteristics of reported residential burglary in Bellevue;

(2) the extent of citizen involvement in the Bellevue burglary prevention project; and

(3) the impact of the burglary prevention program on the reported

residential burglary rates for Bellevue.

In summary, the study showed that:

(1) Bellevue as a whole experienced a significant reduction in both rates and frequency of reported residential burglary; and

(2) the intensive participant neighborhoods (Ardmore and Enatai) experienced a greater reduction in both reported residential burglary rates and frequency than the rest of Bellevue.

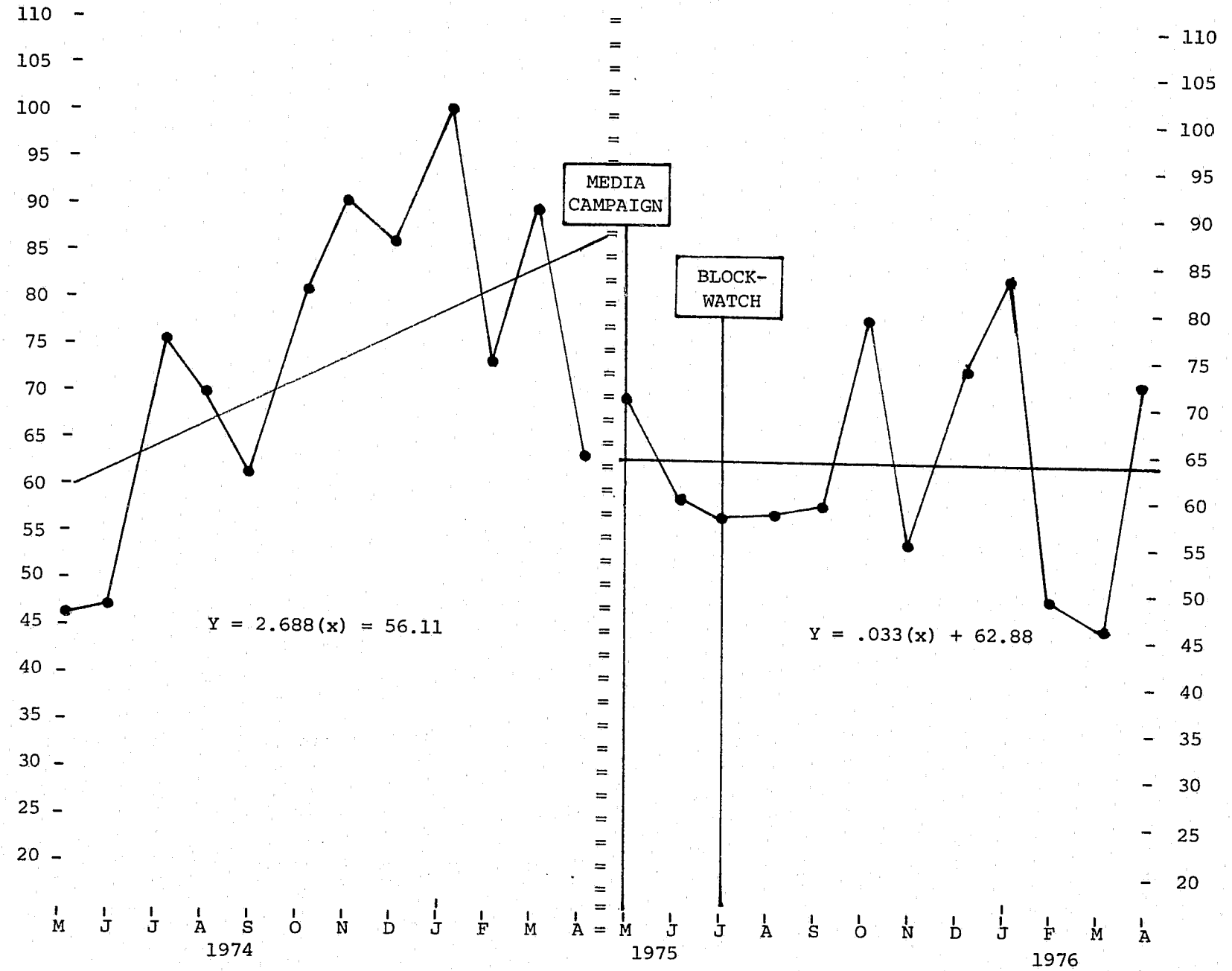
Furthermore, the actual decreases in reported residential burglary frequency and rates may actually be greater than those noted in this evaluation if residential burglary reporting rates increased as a result of the project.

While the decreases in Bellevue's reported residential burglary frequency and rates were significant, it was not possible to attribute the decreases to the burglary prevention program alone. Other factors may have partially accounted for the decreases. Other possible causes which may have contributed to the apparent decreases in reported residential burglary in Bellevue included the effects of statistical regression, and/or the combined effects of Bellevue's Team Policing strategies. Furthermore, it was not known to what extent, if any, burglary displacement effects may have influenced the project's outcome.

The qualifications notwithstanding, the data contained in this evaluation suggest that the program has been a success. It appears that crime prevention activities of this nature are worthwhile and therefore should be continued within the structure of Bellevue's Team Policing Program.

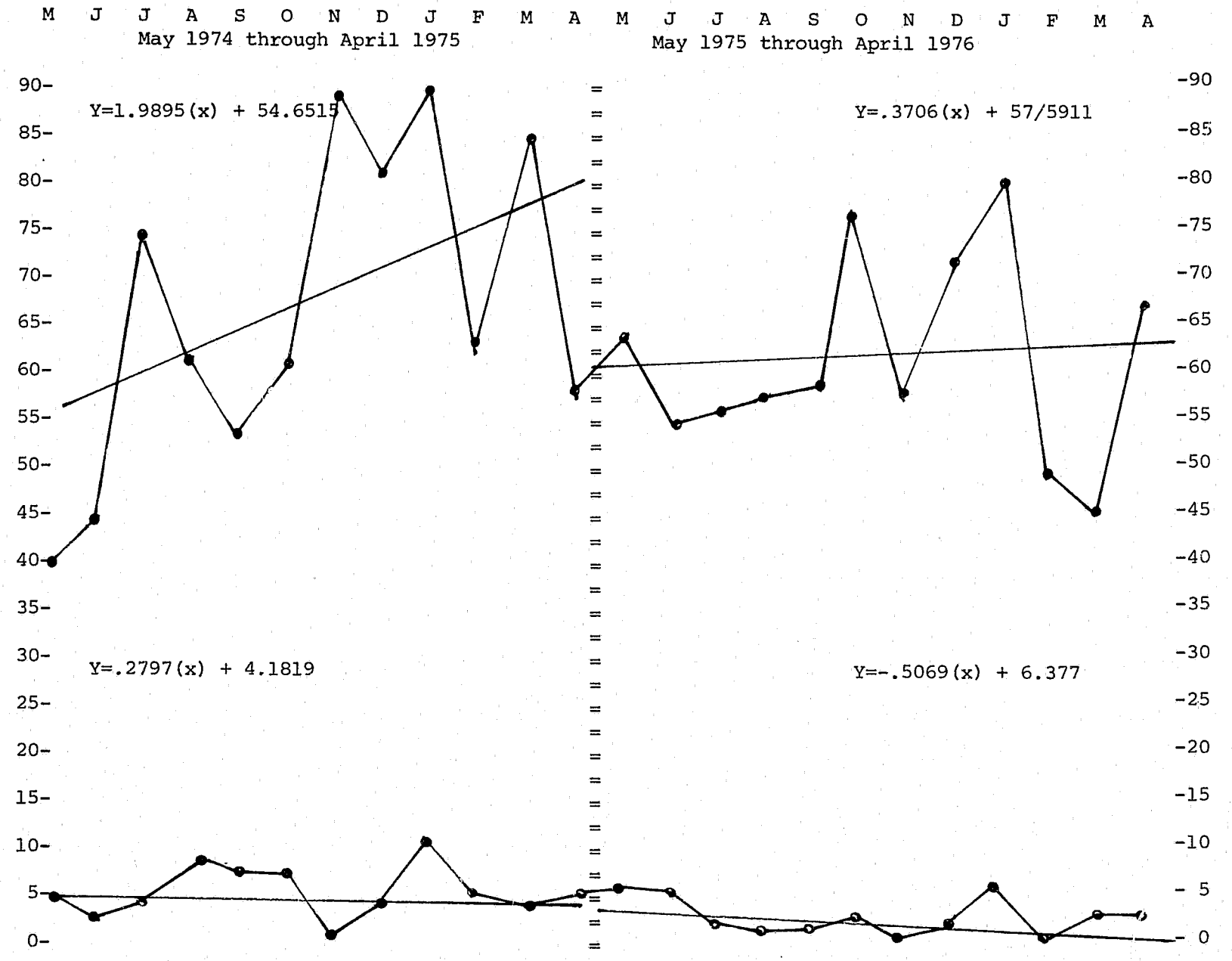
G R A P H 2

TOTAL RESIDENTIAL BURGLARIES -- MAY 1974 - APRIL 1976



G R A P H 3

RESIDENTIAL BURGLARIES 1974/75 - 1975/76



APPENDIX XI

REPORTED RESIDENTIAL BURGLARIES PER MONTH

PRE May 1974 through April 1975

POST May 1975 through April 1976

	1. Ardmore		2. Enatai		3. Total 1&2		4. Bellevue (-3)	
	PRE	POST	PRE	POST	PRE	POST	PRE	POST
May	4	9	1	0	5	9	41	60
June	2	6	1	2	3	8	44	52
July	2	3	2	0	4	3	73	54
August	6	1	2	1	8	2	62	56
September	2	1	5	1	7	2	54	57
October	5	3	2	0	7	3	74	75
November	2	0	0	0	2	0	89	56
December	5	1	0	1	5	2	81	71
January	4	1	6	3	10	4	89	79
February	4	0	3	0	7	0	65	49
March	4	2	2	0	6	2	84	45
April	6	1	2	1	8	2	55	66
Total	46	28	26	9	72	37	811	720
% Change	-39.1%		+65.4%		-48.6%		-11.2%	

Discussion of "Bellevue Citizen Involvement in Burglary Prevention"

Perhaps the greatest strength of Mulberg and Wein's evaluation is their comparison of the neighborhoods targeted for extensive treatment with the rest of Bellevue. The technique offers an estimate of the reduction in burglary which might be expected from expanded use of the extensive program. However, this approach becomes rather complex as it creates four distinct time series (the pre and post series for the targeted neighborhoods and the pre and post series for the rest of Bellevue) which the authors display in their Graph 3. Here the authors report the slope and intercept coefficients for the four time series. Using these descriptive parameters the authors concluded that there was a discernible effect from the extensive application of burglary reduction measures in the targeted neighborhoods. This finding is bolstered by a statistically significant t-test comparing the mean of the pre series with the mean of the post series.

Using the comparison group (the rest of Bellevue) in this manner actually under-utilizes the data presented. For example, one of the major values of a multiple time series analysis is that it allows the researcher to control for trends and instabilities in the data. Visual examination of Graph 3 in the report suggests the presence of trends and particularly instability (sharp fluctuations up and down in the number of reported burglaries) in the comparison group. In order to take advantage of the additional information embodied in the comparison time series, Campbell and Stanley suggest differences between the series can be analyzed: In this case, the differences between the number of burglaries reported in the targeted neighborhoods and the rest of the city. However, the differences in these data also fluctuate widely. A

ratio (burglaries reported in the targeted neighborhoods as a percentage of burglaries in the rest of Bellevue) offers an alternative method to control for instability and trends. These ratios are plotted in Figure 1 and indicate instability is still present. However, new aspects of these data are visible when cast in this form.

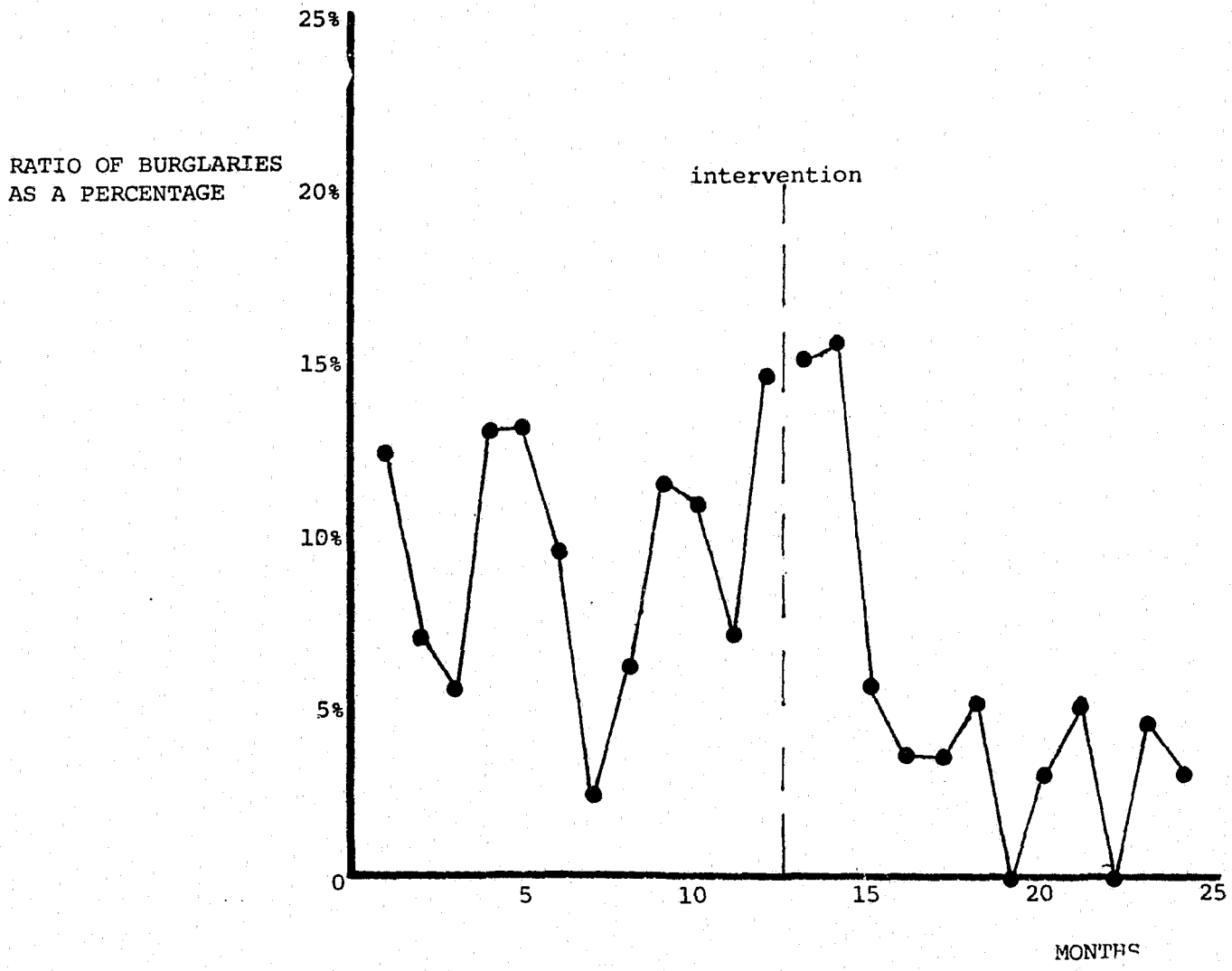
Examining the pre intervention series, it is clear that before the program the target neighborhoods accounted for a random proportion of reported burglaries. In contrast, after the intervention the fluctuations of the proportions are less severe. When the data are presented as in Graph 3 of the report, this change is not evident and consequently went unnoticed by the authors. This is unfortunate because this decrease in fluctuations may be the most important effect of the burglary reduction program. Interpretation of this change depends on additional information and perhaps consultation with the Bellevue police department might provide some understanding. For example, the police might know that they have curtailed the operation of professional burglars in the target neighborhoods, or alternatively they have evidence that they have frightened away local amateurs. In any event, this finding cannot be interpreted without additional information. The point is, however, that unless such a change is observed, it cannot be followed up by the evaluators. In this case, re-analysis of the data by combining the time series for treatment and comparison groups did not remove the instability from these data as hoped, but focused attention on the relative level of fluctuations in the time series.

Perhaps the most important aspect of Figure 1 is that is reveals a startup lag. The first two points in the post intervention series

FIGURE 1

TIME SERIES PLOTS FOR THE RATIO OF BURGLARIES IN TARGETED NEIGHBORHOODS

TO THE REST OF BELLEVUE STATED AS A PERCENTAGE

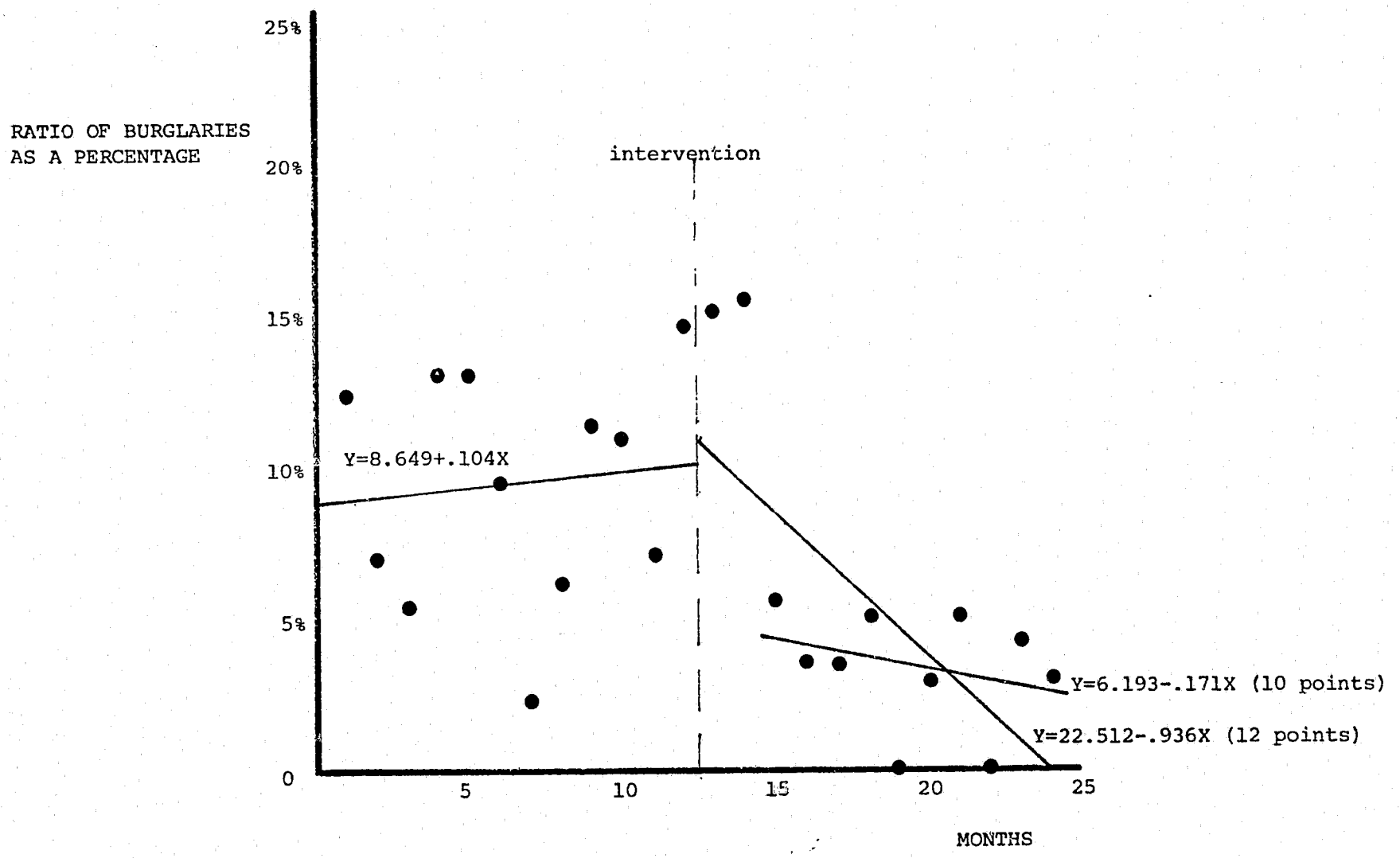


are in fact the highest points in the entire two year period. This suggests the effects of the burglary reduction measures were not immediate but delayed for at least three months. In their narrative discussion Mulberg and Wein recognize this possibility and point out that at the outset relatively few people had been contacted by the program. Examination of their Graph 3 does not suggest a startup lag. Only by combining the two series does this effect become apparent. In this situation it makes little sense to analyze these first two points as part of the post intervention series. Morecver, their inclusion produces distortions that can be very misleading. Figure 2 illustrates the effect of including these two startup points: When they are included the slope of the post series is steeply negative (-.936). This suggests that the targeted neighborhoods are reducing their proportion of burglaries by almost one percent per month. If this were the true effect, the targeted neighborhoods would have reduced their proportion to zero by the end of the first year of operation of the program. Visual examination of the points indicates that such optimism is unwarranted. Removing the two startup points produces a more gradual negative slope (-.171), indicating the extensive application of the program may be having the desired effect but not nearly as effectively as suggested by the 12-point slope.

Recognition of the two startup lag points in the post intervention series also calls attention to the validity of the t-test procedure used by Mulberg and Wein. The t-test assesses the significance of the difference of the pre and post series summarized by their means. Detection of the two startup points suggests we should attempt to remove their effect before we conduct the statistical test. However, rather than

FIGURE 2

THE EFFECT OF STARTUP LAG ON POST INTERVENTION SLOPE ESTIMATES FOR RATIOS



replicate the t-test, deleting the startup points, it is more useful to statistically analyze the entire time series. This involves a comparison of the slopes and intercepts of the pre and post intervention regression lines by means of three separate statistical tests developed by Walker and Lev. These tests can be conducted by making use of the time series program available at the computing centers of the University of Washington, Eastern Washington State College, and Western Washington University. This program is referenced in the "Computer Resources" section of this handbook. The three tests can be informally summarized as follows:

Walker-Lev 1

Null Hypothesis: Separate independent regression lines for the pre and post intervention series do not fit the data points better than two regression lines with a common (the same) slope.

Walker-Lev 2

Null Hypothesis: The common slope of the regression lines (referred to in test 1) is not different from zero.

Walker-Lev 3

Null Hypothesis: Separate regression lines for the pre and post intervention series do not fit the data better than a single regression line.

The logic of inference using these tests is cumulative and may involve more than one of the tests. If test 1 is significant the null hypothesis is rejected and we infer a change has taken place in the slope of the proportion of burglaries after the intervention. This implies that the burglary reduction program has caused the proportion of

burglaries in the targeted neighborhoods to decrease over time. Essentially we would have some evidence that the program is working as desired. However, it is possible that a change may have occurred without altering the slopes of the pre and post series. For example, the proportion of burglaries in the targeted neighborhoods could be reduced but continue to decrease (or increase) at the same rate before and after the intervention. This would be a program effect that changes the intercept of the regression line. Using test 2 it is possible to examine the steepness of this common slope. If test 2 is significant, we conclude the common rate of change in the proportion is significantly decreasing, both before and after the intervention. This can be interpreted as the common trend of the two series. If test 2 is not significant, we infer there is no trend. This implies the proportion of burglaries remains statistically constant before and after the intervention. If test 3 is significant, we infer that the pre intervention series is statistically independent of the post intervention series. Taken in conjunction with a finding of no significance in test 1, this implies a significant step level change has taken place after the intervention. If test 3 is significant then we would conclude that a single line fits both the pre and post intervention points and that there is no evidence of an effect from the burglary reduction program.

Applying these tests to the ratio data displayed in Figure 1 and 2, it is possible to get a more complete understanding of these time series and still apply rigorous statistical tests with which we can make inferences about program effectiveness. The results of the three Walker-Lev tests for the ratios are shown in Table 1. These tables are excerpted photocopies from the printout of the interactive version of the time

TABLE 1

REGRESSION DISCONTINUITY STATISTICS FOR THE RATIO OF BURGLARIES IN TARGET NEIGHBORHOODS TO BURGLARIES IN THE REST OF BELLEVUE STATED AS A PERCENTAGE

Twelve (12) Post-Intervention Points

WALKER-LEV 1 TEST

F-RATIO = 5.191 WITH 1, 20 DEGREES OF FREEDOM p<.05
SEPARATE GPS: PRE-X GP PREDICTED Y = 9.955
SLOPE = 0.104 INTCPT = 8.649
POST-X GP PREDICTED Y = 10.813
SLOPE = -0.936 INTCPT = 22.512
WITHIN GPS: PRE-X GP PREDICTED Y = 6.834
SLOPE = -0.416 INTCPT = 12.030
POST-X GP PREDICTED Y = 7.692
SLOPE = -0.416 INTCPT = 12.888

WALKER-LEV 2 TEST

F-RATIO = 2.632 WITH 1, 20 DEGREES OF FREEDOM not significant

WALKER-LEV 3 TEST

F-RATIO = 0.061 WITH 1, 21 DEGREES OF FREEDOM not significant
COMBINED GPS: PREDICTED Y = 7.263
SLOPE = -0.362 INTCPT = 11.788

Ten (10) Post-Intervention Points

WALKER-LEV 1 TEST

F-RATIO = 0.394 WITH 1, 18 DEGREES OF FREEDOM not significant
SEPARATE GPS: PRE-X GP PREDICTED Y = 9.955
SLOPE = 0.104 INTCPT = 8.649
POST-X GP PREDICTED Y = 4.054
SLOPE = -0.171 INTCPT = 6.193
WITHIN GPS: PRE-X GP PREDICTED Y = 9.350
SLOPE = 0.004 INTCPT = 9.305
POST-X GP PREDICTED Y = 3.181
SLOPE = 0.004 INTCPT = 3.135

WALKER-LEV 2 TEST

F-RATIO = 0.000 WITH 1, 18 DEGREES OF FREEDOM not significant

WALKER-LEV 3 TEST

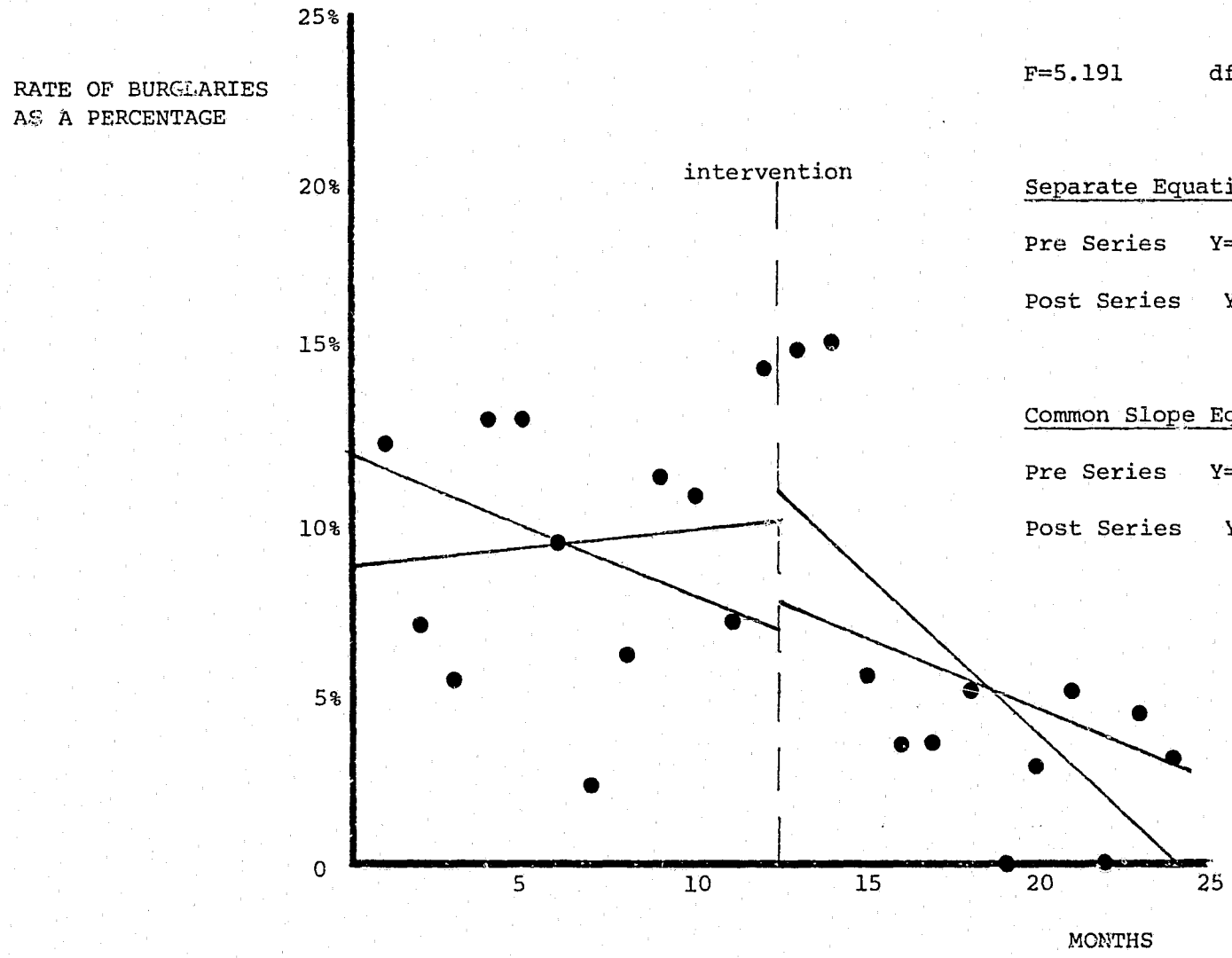
F-RATIO = 5.422 WITH 1, 19 DEGREES OF FREEDOM p<.05
COMBINED GPS: PREDICTED Y = 6.128
SLOPE = -0.414 INTCPT = 11.308

series program available at the University of Washington. Other output is produced by this program but has been omitted here for simplicity. The raw printout has been copied to give the reader experience with the actual form of the results which are produced by the computer program. The top half of Table 1 includes the two startup points. The bottom half excludes them. When the two points are included, test 1 is significant (F=5.191), while tests 2 and 3 are not significant. This implies that a change of slope has occurred after the intervention. When the two startup points are removed, only test 3 is significant, which implies that the series have a common slope (test 1) which is not different from zero (test 2) that cannot be fitted with a single line (test 3). Taken together these three tests imply that a significant change has taken place but that it is a step-level change or a change in intercepts. Moreover, test 2 implies that the rate of change in the common slopes is in fact statistically constant. Figures 3 and 4 are included here to display the various slopes and intercepts listed in the three tests. Visual inspection makes it clear that the twelve point post intervention series (including the startup points) leads to rather different conclusions than implied by the ten point post intervention series. In Figure 3-B it is easy to see, for example, that the common slopes based on the ten point series are virtually flat. In contrast, the distortions induced by these points (shown in Figure 3-A) suggest more drastic change.

An important aspect of these data that needs to be re-emphasized in the context of these statistical tests is the extreme variability or instability of the pre intervention series. Because these points fluctuate so widely, any test of significance, whether it is based on means or on intercepts and slopes is not a "good" test. In the case of means

FIGURE 3-A

WALKER-LEV TEST 1 FOR 12 POST INTERVENTION POINTS



F=5.191 df=1,20 p<.05

Separate Equations

Pre Series Y=8.649+.104X

Post Series Y=22.512-.936X

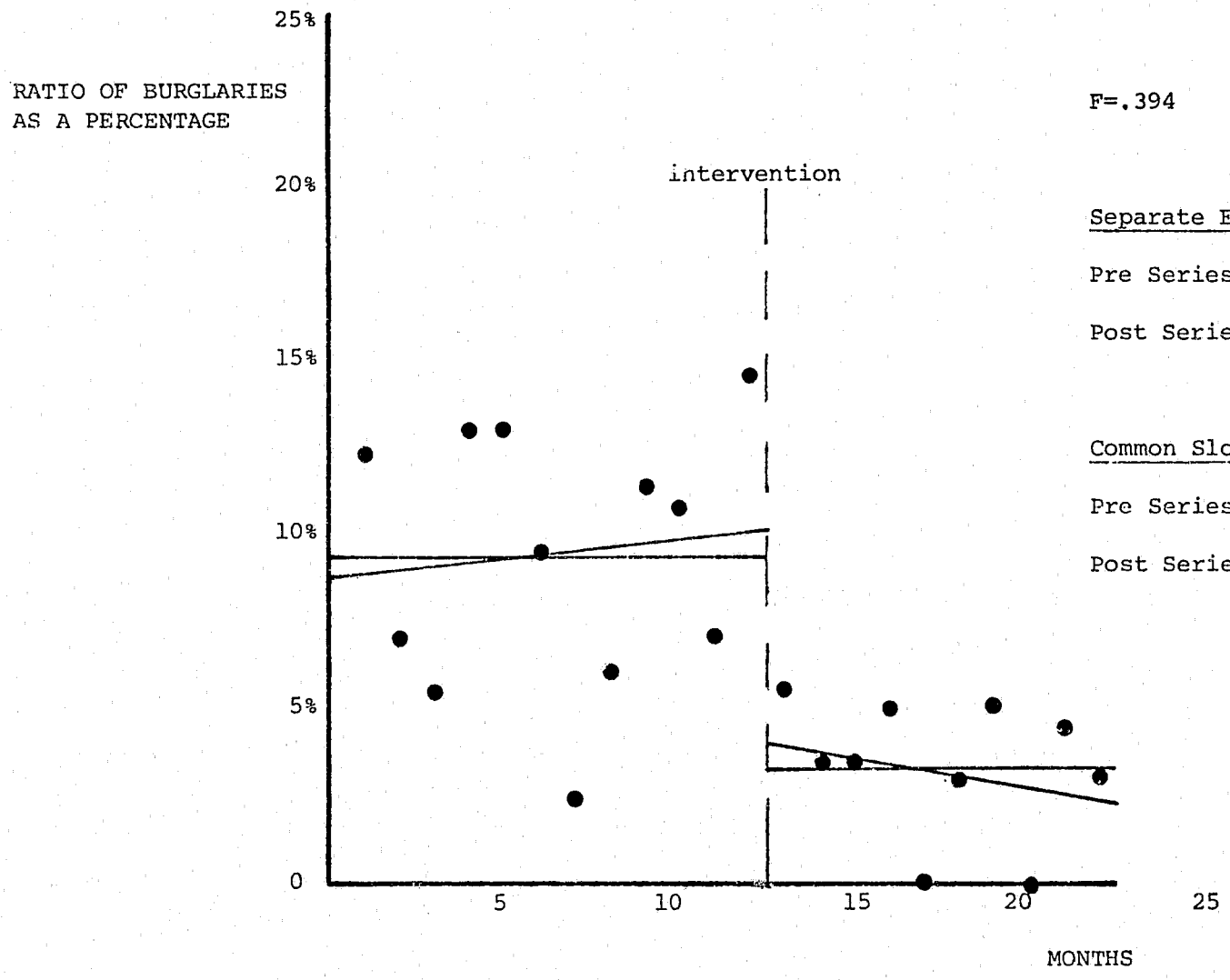
Common Slope Equations

Pre Series Y=12.03-.416X

Post Series Y=12.888-.416X

FIGURE 3-B

WALKER LEV TEST 1 FOR 10 POST INTERVENTION POINTS



F=.394 df=1,18 N.S.

Separate Equations

Pre Series Y=8.649+.104X

Post Series Y=6.193-.171X

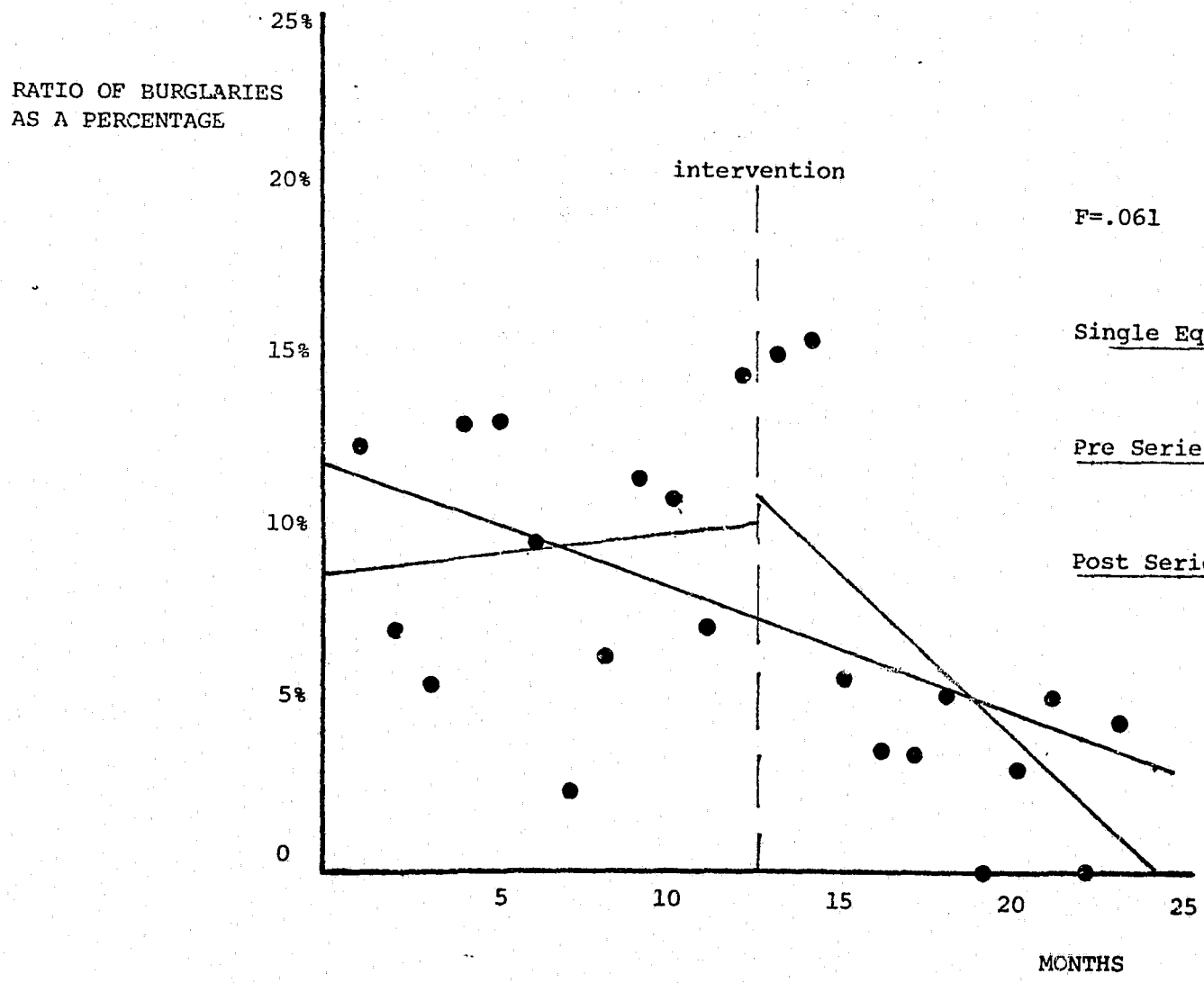
Common Slope Equations

Pre Series Y=9.406+.004X

Post Series Y=3.135+.004X

FIGURE 4-A

WALKER LEV TEST 3 FOR 12 POST INTERVENTION POINTS



F=.061 df=1, 20 N.S.

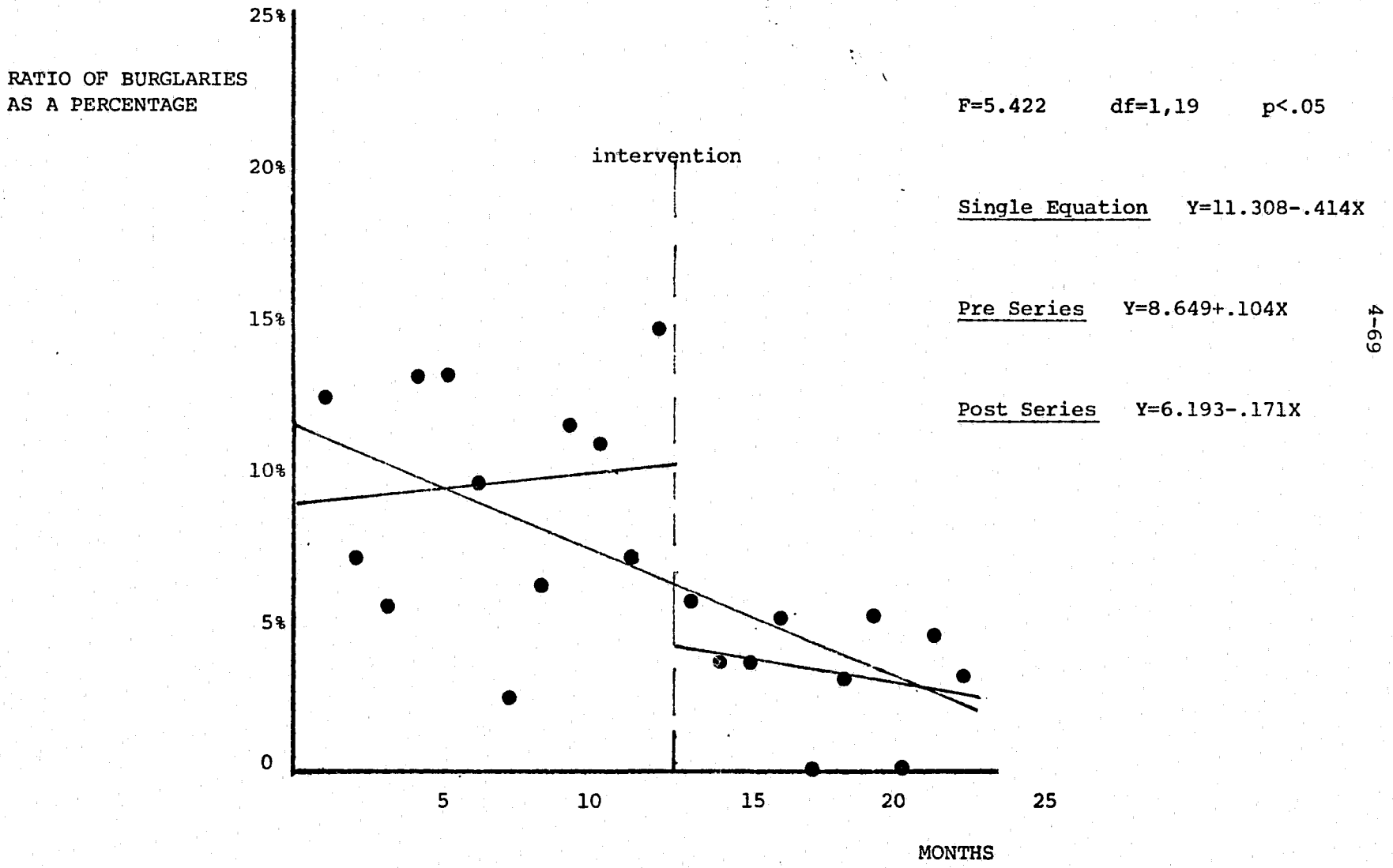
Single Equation Y=11.788-.362X

Pre Series Y=8.649+.104X

Post Series Y=22.512-.936X

FIGURE 4-B

WALKER LEV TEST 3 FOR 10 POST INTERVENTION POINTS



as employed in the t-test by Mulberg and Wein or in the case of regression lines discussed here most of the points are quite divergent from the mean, or the regression lines. Simply put, these measures do not represent the data well because they seek to simplify a set of data that are not orderly and not subject to summary description by any means. Hence the tests suggested here are greatly affected by the apparent randomness of the pre series and their interpretation must be made while recognizing the inherent randomness of these data. Because the pre intervention series is so random, any number of lines might be fitted to these points without greatly reducing the goodness of fit. Since the Walker-Lev tests are based on "improvements" in the goodness of fit, only slight regularities in the post intervention series are capable of producing significant statistics.

Stated in another manner, blind testing of data can be very misleading and only common sense can counter the blindness of statistical procedures.

This re-analysis of Mulberg and Wein's data tends to confirm their findings: The extensive application of the burglary reduction program seems to be working in the target neighborhoods. However, findings from this more textured time series analysis provide additional insights. There is evidence of a short startup lag, there seems to be no significant trend in the proportion of burglaries in the targeted neighborhoods, but there is evidence that a step-level change has taken place such that the proportion of burglaries in the targeted area has been reduced. These additional findings are all based on ratios rather than raw numbers of burglaries. However, a parallel analysis of the raw figures for both the target area and the rest of Bellevue is provided here (Tables 2 and

TABLE 2

REGRESSION DISCONTINUITY STATISTICS FOR THE NUMBER OF BURGLARIES IN THE NEIGHBORHOODS TARGETED FOR EXTENSIVE TREATMENT

Twelve (12) Post-Intervention Points

WALKER-LEV 1 TEST

F-RATIO = 9.319 WITH 1, 20 DEGREES OF FREEDOM p<.01
SEPARATE GPS: PRE-X GP PREDICTED Y = 7.678
SLOPE = 0.280 INTCPT = 4.182
POST-X GP PREDICTED Y = 6.125
SLOPE = -0.507 INTCPT = 12.463
WITHIN GPS: PRE-X GP PREDICTED Y = 5.318
SLOPE = -0.114 INTCPT = 6.739
POST-X GP PREDICTED Y = 3.765
SLOPE = -0.114 INTCPT = 5.186

WALKER-LEV 2 TEST

F-RATIO = 0.531 WITH 1, 20 DEGREES OF FREEDOM not significant

WALKER-LEV 3 TEST

F-RATIO = 0.543 WITH 1, 21 DEGREES OF FREEDOM not significant
COMBINED GPS: PREDICTED Y = 4.542
SLOPE = -0.211 INTCPT = 7.178

Ten (10) Post-Intervention Points

WALKER-LEV 1 TEST

F-RATIO = 2.079 WITH 1, 18 DEGREES OF FREEDOM not significant
SEPARATE GPS: PRE-X GP PREDICTED Y = 7.678
SLOPE = 0.280 INTCPT = 4.182
POST-X GP PREDICTED Y = 2.424
SLOPE = -0.085 INTCPT = 3.485
WITHIN GPS: PRE-X GP PREDICTED Y = 6.878
SLOPE = 0.146 INTCPT = 5.049
POST-X GP PREDICTED Y = 1.268
SLOPE = 0.146 INTCPT = -0.561

WALKER-LEV 2 TEST

F-RATIO = 1.294 WITH 1, 18 DEGREES OF FREEDOM not significant

WALKER-LEV 3 TEST

F-RATIO = 12.365 WITH 1, 19 DEGREES OF FREEDOM p<.01
COMBINED GPS: PREDICTED Y = 3.948
SLOPE = -0.234 INTCPT = 6.870

TABLE 3

REGRESSION DISCONTINUITY STATISTICS

FOR THE NUMBER OF BURGLARIES IN THE CITY OF BELLEVUE

EXCLUDING NEIGHBORHOODS TARGETED FOR EXTENSIVE TREATMENT

Twelve (12) Post-Intervention Points

WALKER-LEV 1 TEST

F-RATIO = 1.717 WITH 1, 20 DEGREES OF FREEDOM not significant
SEPARATE GPS: PRE-X GP PREDICTED Y = 82.038
SLOPE = 2.409 INTCPT = 51.924
POST-X GP PREDICTED Y = 57.776
SLOPE = 0.371 INTCPT = 53.143
WITHIN GPS: PRE-X GP PREDICTED Y = 75.923
SLOPE = 1.390 INTCPT = 58.549
POST-X GP PREDICTED Y = 51.661
SLOPE = 1.390 INTCPT = 34.288

WALKER-LEV 2 TEST

F-RATIO = 2.940 WITH 1, 20 DEGREES OF FREEDOM not significant

WALKER-LEV 3 TEST

F-RATIO = 4.908 WITH 1, 21 DEGREES OF FREEDOM p<.05
COMBINED GPS: PREDICTED Y = 63.792
SLOPE = -0.129 INTCPT = 65.406

WALKER-LEV 1 TEST

F-RATIO = 1.461 WITH 1, 18 DEGREES OF FREEDOM not significant
SEPARATE GPS: PRE-X GP PREDICTED Y = 82.038
SLOPE = 2.409 INTCPT = 51.924
POST-X GP PREDICTED Y = 60.255
SLOPE = .109 INTCPT = 58.891
WITHIN GPS: PRE-X GP PREDICTED Y = 76.989
SLOPE = 1.568 INTCPT = 57.394
POST-X GP PREDICTED Y = 52.962
SLOPE = 1.568 INTCPT = 33.367

WALKER-LEV 2 TEST

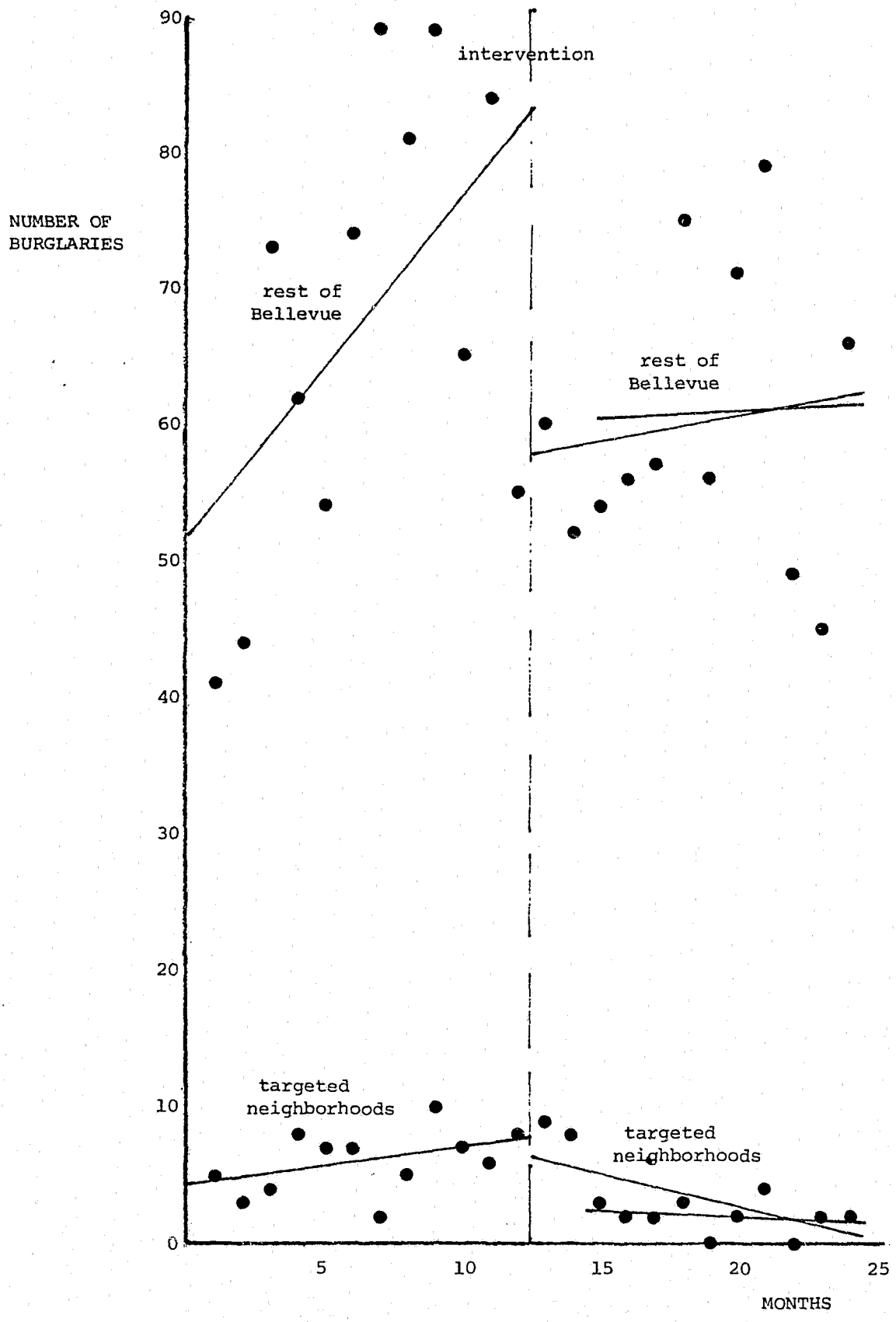
F-RATIO = 2.706 WITH 1, 18 DEGREES OF FREEDOM not significant

WALKER-LEV 3 TEST

F-RATIO = 4.133 WITH 1, 19 DEGREES OF FREEDOM not significant
COMBINED GPS: PREDICTED Y = 64.440
SLOPE = -.060 INTCPT = 65.195

FIGURE 5

STARTUP LAG EFFECTS ON POST INTERVENTION SLOPES FOR NUMBER OF BURGLARIES



3 contain the Walker-Lev statistics and Figure 5 illustrates the differences induced by the startup points in the raw data). The reader is urged to study and compare these analyses as an illustrative example of the types of problems likely to be encountered in an actual regression discontinuity analysis.

CONTINUED

4 OF 7

SECTION 4.C.

CLARK COUNTY (VANCOUVER, WASHINGTON) DEINSTITUTIONALIZATION OF STATUS OFFENDERS PROJECT EVALUATION REPORT*

Office of Juvenile Justice and Delinquency Prevention
and
Vancouver (Clark County), Washington, Juvenile Court

By

Anne L. Schneider
Institute of Policy Analysis

August, 1978

*Selections from the Clark County (Vancouver, Washington) Deinstitutionalization of Status Offenders Project Evaluation Report have been excerpted for inclusion.

I N T R O D U C T O R Y C O M M E N T S O N :

"CLARK COUNTY (VANCOUVER, WASHINGTON) DEINSTITUTIONALIZATION OF STATUS OFFENDERS PROJECT EVALUATION REPORT"

Projects that seek to reduce recidivism rates are among the most difficult types to evaluate because of problems in measuring recidivism and in controlling for alternative explanations. In the Vancouver DSO project, a random assignment of eligible DSO youths to experimental and control conditions was undertaken, but became somewhat confounded for reasons that were never entirely clear to the evaluators. Even though the control and experimental groups were quite similar on most relevant characteristics, the evaluators approached the analysis as if it were a quasi-experimental rather than experimental design. The techniques illustrated in the report include a time series analysis of individual-level data (using multiple regression) in which all pre-project status offenders are compared to all post-DSO status offenders even though only part of the latter were actually in the DSO project. This technique maximizes equivalency of the pre and post cases. In addition, time series tests were made to rule out the possibility that the project produced changes in the characteristics of status offenders which could be confused with the effect of the project itself.

Multiple regression analysis is used to control for differences between the experimental and control groups. This analysis is substantiated with contingency tables that examine whether the differences in recidivism between experimental and control are consistently in the same direction within various categories of youths.

A second part of the evaluation is excerpted to illustrate a common

problem in time series designs: determining when the project started. The evaluators identify two statistically significant shifts in the proportion of status offenders detained at the juvenile court: one when the local judges approved, in principle, of the application for federal DSO funds, and a second when the DSO project was implemented. A double intervention analysis cannot be done with the Walker-Lev statistical program, but can be accomplished with multiple regression and analysis of covariance.

The multiple regression formula for a double intervention is:

$$Y = a + b_1 I_1 + b_2 \text{Time} + b_3 I_1 \text{ Time} + b_4 I_2 + b_5 I_2 \text{ Time}$$

where: Y = dependent variable

I_1 = one intervention (o = pre I_1, 1 = post I_1)

Time = time, measured 1, 2, 3...n (Months, weeks, years, etc.)

I_1 Time - interaction term (time multipled by I_1)

I_2 - second intervention (0 = pre I_2; 1 = post I_2)

I_2 time - interaction term (time multipled by I_2)

CLARK COUNTY (VANCOUVER, WASHINGTON)
DEINSTITUTIONALIZATION OF STATUS OFFENDERS PROJECT
EVALUATION REPORT

INTRODUCTION

With a $50,000 two-year grant from the Office of Juvenile Justice and Delinquency Prevention (OJJDP), the Vancouver (Clark County), Washington juvenile court began a program to deinstitutionalize status offenders (DSO) in July 1976. The Vancouver project was the smallest of the national DSO grants and most of the funds were used for direct service delivery. The major components of the project were crisis intervention counseling provided by two newly-hired juvenile court probation officers and family crisis intervention counseling provided by volunteers trained and directed by the project probation officers. The objects of the program were to:

1. Reduce the penetration of status offenders into the juvenile court system by reducing the number detained, reducing commitments for incarceration to the Department of Social and Health Services, and reducing the number of status offenders on whom formal petitions were filed; and

2. Reduce the recidivism of status offenders.

DESCRIPTION OF THE PROJECT

The Vancouver DSO program is operated as a part of the probation unit of the juvenile court. Prior to implementing the deinstitutionalization project, the common practice was for status offenders to be held in detention before being seen by a probation officer, and they were sometimes held in detention

for several days after that time awaiting a counselor from the Department of Social and Health Services (DSHS). The two additional probation officers hired with the federal funds counsel status offenders immediately after court intake in an effort to return them to their homes or to find community-based alternatives to detention. A second component of the DSO program in Vancouver is the development of a group of volunteers who, under the guidance of a probation officer, can provide family crisis counseling. The goal of this portion of the DSO program is to return youths to their homes, thereby making available the extremely limited community bedspace to other youths who are unwilling or unable to return to their homes. In conjunction with DSHS, the Vancouver juvenile court has been attempting to increase the availability of community-based alternatives for status offenders who cannot (or will not) return home. This effort has resulted in twelve additional overnight places reserved for status offenders. The total number of places (other than detention) for short-term care of all juveniles is 78; twelve of these are reserved exclusively for status offenders.

At the time the DSO counselors were hired, the two probation officers who had previously been responsible for status offenders retained their responsibilities by providing counseling to status offenders who were not eligible for the DSO project and those who were in the control group. Thus, the open case load for status offender probation officers was reduced simultaneously with the implementation of the project.

During the time that the DSO project was operative in Clark County, the juvenile court system had several key decision points that could result in the case being continued on through court processing or terminated. A flow chart of the court procedures, a description of who did what, the criteria upon which decisions were based, and an analysis of the number of cases flowing through various parts of the system are contained in Appendix A. In general,

status offenders could be referred to the court from eight different law enforcement agencies, schools, parents, and other jurisdictions. The referrals could be in person (e.g., the youth appears at court intake), or they could be paper referrals. For the personal referrals, the court intake officer conducted an initial screening of the case and, if a probation officer was available to talk with the youth and/or family, the case would be referred immediately to probation. The probation officer, in this situation, could determine whether a detention hearing would be needed and had three options for disposal of the case: (1) informal adjustment whereby the youth and probation officer reached agreement concerning the youth's activities (this normally involved no followup or only very limited followup by the probation officer); (2) informal probation whereby the parents, youth, and probation officer reached agreement on the youth's activities (this normally was accompanied by limited followup); and (3) the filing of a status offense petition against the youth, which would be followed by a fact-finding and disposition hearing.

In the event that no probation officer was available to talk with the youth at intake, the intake officer would determine whether the youth should be detained or not and, if the youth was not to be detained, he or she was asked to return the next day (or within a few days) to talk with a probation officer. If the youth was detained, an appointment with a probation officer would be made for the next day.

Paper referrals to the court on status offenders were sent directly to the head of the status offender probation unit. The probation officer would then attempt to contact the youth and family involved in the offense. If contact was made, an appointment would be set for the youth and family to discuss the situation with a probation officer. Not all paper referrals, of course, resulted in contact with anyone at the court.

With the implementation of the crisis intervention DSO project, it was expected that the number of status offenders detained would decline because of the fact that the DSO counselors would be on duty for weekends and for longer hours during the week (8:00 AM through 11:00 PM) rather than the normal daytime shift, and because of their efforts to be available for immediate counseling of the youth and family rather than having their calendars full of prescheduled appointments. The crisis intervention counseling, family counseling, and decline in detention were expected to reduce the need to file petitions against the youths because they expected to be able to resolve a larger proportion of the disputes, enabling the youths to return home or to an acceptable community alternative.

Incarceration in Clark County was not, technically, done by the juvenile court. Rather, the court could commit status offenders to the Department of Social and Health Services (DSHS), with the stipulation that the youth needed foster care or with the stipulation that the youth might need to be institutionalized. DSHS made the final decision on this.

The reduction in recidivism of status offenders was expected to result from the reduced penetration of the youths into the system and/or to the nature of the counseling. Underlying the expectation that reduced penetration would in turn reduce recidivism is the idea that youths who come into contact with the juvenile court and who remain in contact with it for a longer period of time are labelled by themselves and others as problem youths, which tends to produce more problem behavior in subsequent months.[FN] One could argue, from a deterrence perspective, however, that the lack of penalty for running away, curfew violations, truancy, or incorrigible behavior would result in a youth believing that these types of problem behavior would evoke no official penalty and therefore could be continued.[FN]

IMPACT OF THE DSO PROJECT ON DETENTION OF STATUS OFFENDERS

A major purpose of the federal DSO initiative was to prevent status offenders from having to spend time in detention and, hence, to reduce the length of their contact with the juvenile justice system.

In order to determine whether the Clark County project reduced the proportion of status offenders in detention, a statistically significant change should occur from the pre to post time periods and this change must be attributable to DSO rather than to other factors which might have produced it. As noted previously, the random assignment procedure was not imeplemented nor adhered to properly and biases were introduced into the control and experimental groups. Thus, straightforward comparisons of these groups in terms of detention proportions cannot be used to draw conclusions about the effectiveness of DSO. Instead, two types of quasi-experimental procedures will be used to judge the evidence about the effect of the project on detention: interrupted time series analysis of proportion detained per month and a multiple regression prediction technique that will statistically adjust for differences attributable to factors other than the project in order to isolate the independent impact of DSO on detention.

Change in the Pre-Post Detention Patterns

Figures 5 and 6 (and Appendix B) contain the information from the time series analysis of detention. Several observations can be made:

First, the proportion of all status offenders who were detained in juvenile hall increased rapidly from January 1974 to circa July 1975, with the average being approximately 2.6 percent more of the status offenders detained per month (see Figure 5). At this point, a statistically

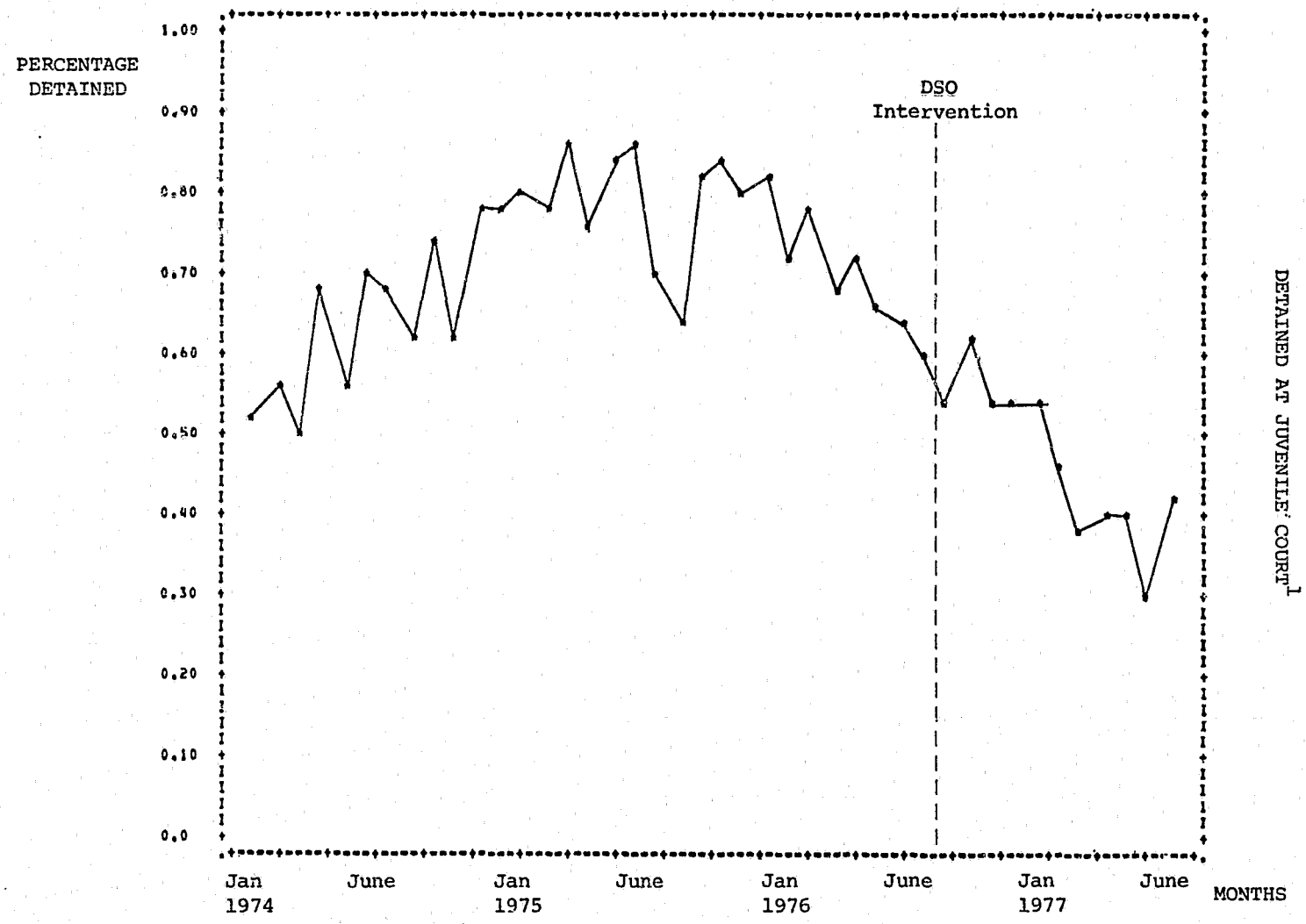


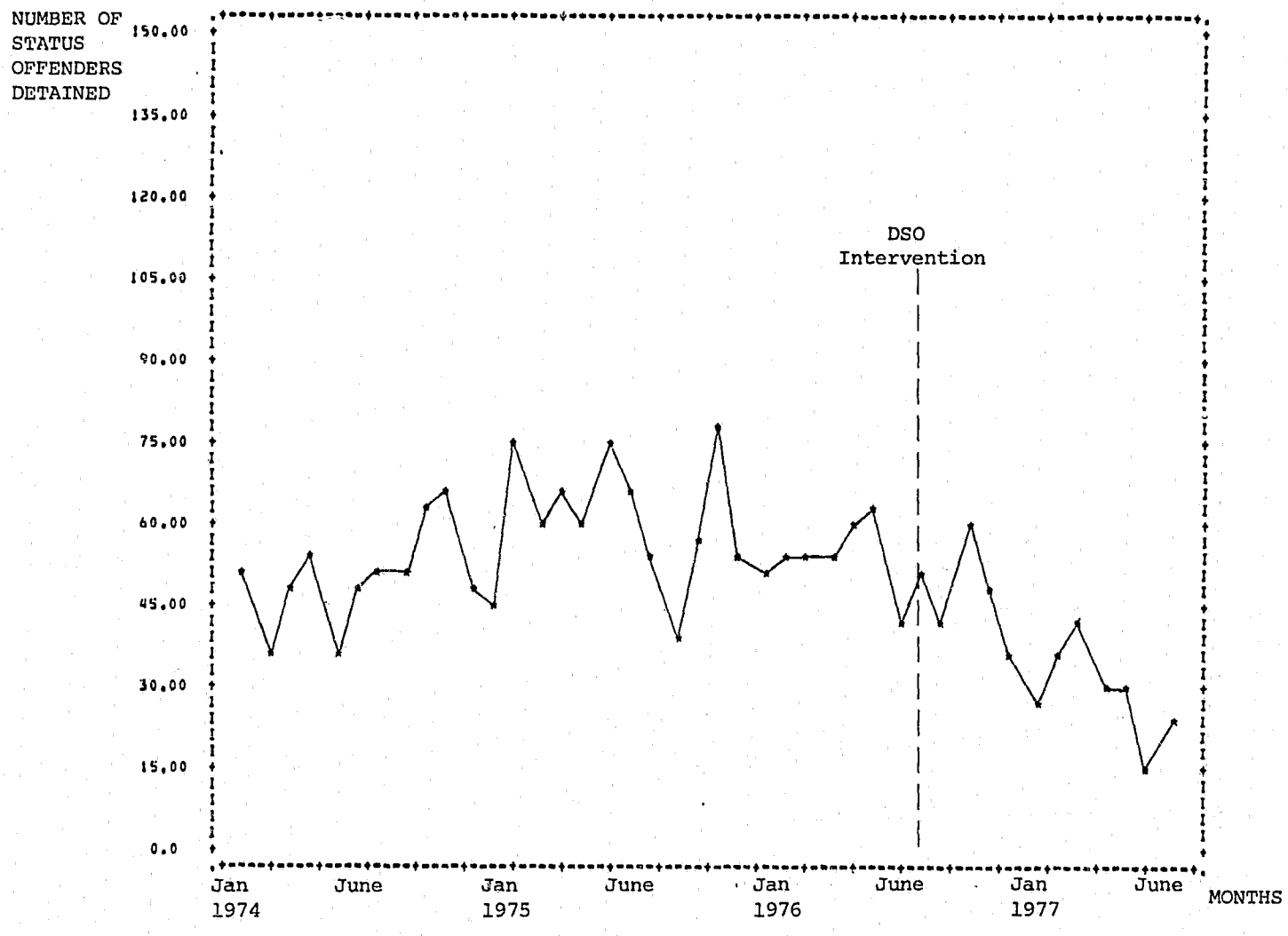
FIGURE 5

PERCENTAGE OF STATUS OFFENDER REFERRALS DETAINED AT JUVENILE COURT[1]

[1]The results (see Appendix B) show that a statistically significant change occurred in the summer of 1975 when the DSO application was approved (locally) and when DSO was implemented in July 1976.

FIGURE 6

NUMBER OF STATUS OFFENDER REFERRALS DETAINED AT JUVENILE COURT[1]



[1]The results (see Appendix B) show that a statistically significant change occurred in the summer of 1975 when the DSO application was approved (locally) and when DSO was implemented in July 1976.

significant change in detention occurred. From July 1975, the proportion detained declined steadily at a rate of about .74 percent of the total status offenders per month. When the DSO project began in July 1976, an additional decline (significant beyond the .05 level) in the proportion detained is observed. (The post DSO data shown in Figure 5 include all status offenders at the court: experimental, control, and ineligible.)

Second, the actual number of status offenders detained shows a similar pattern (see Figure 6). There is an increase from January 1974 through about July 1975, followed by a decrease that apparently is accellerated when the DSO project began in July 1976.

These results suggest the need to identify the event or change around July 1975 that produced the significant downturn in the percent of status offenders being detained.

The change in proportion of total status offender referrals detained could be explained either by a change in the criteria used in making detention decisions or by a change during the summer of 1975 in the characteristics of status offenders such that detention would be appropriate for a smaller proportion of the referrals.

Analysis of status offender characteristics, discussed previously, indicates no change of the type observed in Figures 5 and 6 in any of the social, economic, or demographic characteristics of the status offenders. Cross-offender regressions on status offender referrals indicate that detention decisions are significantly related to parental status, age, month of referral, introduction of the DSO program, sex, the total number of prior status offenses, the total number of prior delinquent offenses, the total number of all types of offenses combined, and the type of offense for which

the referral was made. Even so, all these variables together account for only 11 percent of the variance in detention decisions. Therefore, even if there had been changes in the criteria used in making detention decisions (rather than a change in the general policy about detention of status offenders), the shift would not have accounted for the marked downturn in the proportion detained that occurred in the summer of 1975.

It is more likely that some exogenous event produced the mid-summer change in detention proportion during 1975. Information from the Vancouver court is that there was no legislative change in the summer of 1975 that could have had any impact on the court (including House Bill 371). Bob Axlund, court administrator, noted that the application for the DSO grant was being considered in the summer of 1975 by the juvenile judges and key members of the court staff. It was during this time period that key personnel, including the judges, agreed to support an application for funds under the federal DSO initiative. It appears that the most likely explanation for the obvious shift in status offender detention rates that occurred in mid-summer 1975 is that it was produced by the anticipation of participating in the federal initiative. This suggests that when court staff and judges became sensitive to the issues of labelling and the plight of the status offenders, they began immediately to respond with actions that were desired by the national program itself. In this sense, the OJJDP initiative might have served as a "consciousness raising" experience for those having contact with status offenders. Although the anticipation of changes that would occur in July 1976 (if the grant were received) apparently prompted a change in court policy about detention of status offenders, there is no way to know whether the downward trend would have continued if the grant had not been awarded.

IMPACT OF DSO ON RECIDIVISM RATES OF STATUS OFFENDERS

The major question to be discussed in this section is whether the DSO intervention brought about a change in the recidivism rates of status offenders. The project could result in reduced recidivism if it is the case, as labelling theorists believe, that youths who experience less penetration into the juvenile justice system are less likely to recidivate. Or the project could effect recidivism independent of its impact on detention and petitions.

As has been done in the previous sections, the analysis will proceed by first examining the impact of the DSO intervention on all status offenders (experimental, control, and ineligible) in order to test the effectiveness of the project on the entire system. In addition, since the post-DSO status offenders are relatively comparable to the pre-DSO youths who committed similar offenses, this provides some assurance that observed differences are not due to changes in the characteristics of the youths. Following these analyses, a comparison will be made between the experimental DSO and control youths in order to ascertain whether the experimental strategy in handling status offenders was more effective, in terms of recidivism, than the control strategy, for youths eligible for the program.

Measurement of Recidivism

Recidivism has been measured in terms of recontact with the juvenile court for either a status or delinquent offense. There are several problems in measurement of recidivism, some of which will be discussed below along with the procedure used in this report to deal with them.

1. The purpose of the DSO project was not simply to reduce the number of subsequent court contacts, but also to reduce the frequency of commission of offenses. And, since youths often commit status or delinquent offenses without being caught or referred to the court, the recontact measure is an underestimate of the actual number of offenses committed.

We have no reason to believe, however, that the proportion of youths referred to the court differed between the experimental and control groups or differed from the pre to post time periods. Thus, even though the recontact measure contains considerable error, the nature of the error is the same for the pre and post time periods and for the experimental, control, and ineligible groups within the post time period. Thus, the major effect of this type of error is that the tests of significance will tend to underestimate the true differences between pre and post, as well as the true differences between experimental and control groups.

2. The number of youths referred to the court for a subsequent offense depends on the number of months the youths were "at risk" after the instant offense. The pre-program youths had far more months in which to commit a subsequent offense than the post-program group. In addition, since the probation officers who handled the control group discontinued their work with eligible status offenders in February 1978, the control group has more months "at risk" than does the experimental group. The best solution to this problem is to select a specific followup time (such as three or six months from the end of the month in which the instant offense was committed). Any instant offenses for which there were not enough months at risk to meet the followup time (three months or six months) are removed from the analysis. This procedure was used here and most of the analysis was based on a three-month followup period. Because data collection ended after the first 12 months of the project, there is a severe reduction of cases in the post period when six or more months of followup data are included.

3. Another problem is what to do with offenses that were committed after the followup time period. One solution is to place the youth who committed the instant offense into the "non-recidivism" category if s/he committed a subsequent offense but it was <u>after</u> the fixed risk period of

three (or six) months. The problem with this approach is that it places youths who we know are going to recidivate into the non-recidivist category and this category already contains many youths who eventually will recidivate. This is particularly true of the pre-program group, in comparison with the post, since the former had longer times at risk. This approach will yield a conservative estimate (underestimate) of the effect of the project unless the full impact of the project occurs during the fixed at risk" time and the project youths do not differ from the others in terms of the proportion recidivating after the fixed risk time. Nevertheless, in the subsequent analysis those persons recidivating after the fixed risk time are counted as non-recidivators.

Change from Pre to Post

One method of assessing the impact of the DSO intervention on the recidivism rates of post-DSO status offenders is to examine the proportion of status offenders (pre and post) who had a subsequent delinquent or status offense within the same month as the instant offense, within two months of the instant offense, within three months of the instant offense, and so on. The results of this analysis are shown in Table 14.

Data in the first row include all of the pre and post cases (since all of them had at least a followup period that extended to the end of the same month in which the instant offense occurred). Within thatmonth, 9 percent of the pre-program status offenders had a subsequent offense compared with 6.3 percent of the post-program status offenders. This difference is significant beyond the .01 level (Z test for significance in proportions). The third column of Table 14 shows the difference between pre and post and the last two columns show the number of cases upon which the analysis is based.

Examination of the first two columns of Table 14 shows that the

TABLE 14

PROPORTION OF STATUS OFFENDERS RECIDIVATING

WITHIN SPECIFIED FOLLOWUP PERIODS, PRE & POST[1]

Number of Months "At Risk"	% with subsequent offense within risk period		Z Value	Difference Pre/Post	Number of Cases[3]	
	Pre	Post[2]			Pre	Post
0 MONTH (same month)	9%	6.3%	2.53	2.7%	2,330	914
1 MONTH	18.9%	16.6%	1.49	2.3%	2,330	860
2 MONTHS	26.8%	21.9%	2.75	4.9%	2.330	807
3 MONTHS	33.1%	25.2%	4.02	7.9%	2,330	729
4 MONTHS	37.2%	29.9%	3.44	7.3%	2,330	651
5 MONTHS	40.1%	32.6%	3.22	7.5%	2,330	542
6 MONTHS	43.7%	35.0%	3.46	8.7%	2,330	465
7 MONTHS	45.7%	37.9%	2.94	7.8%	2,330	416
8 MONTHS	47.6%	39.8%	2.72	7.8%	2,330	349

[1]Recidivism is measured as a subsequent court contact for a delinquent or status offense after the instant status offense. Those who had no subsequent offense within the risk period shown on the left are included as "non-recidivators" when calculating the percentage. The percentages are cumulative across the risk periods. Thus, 18.9 percent of pre-program status offenders had a subsequent offense during the same month or within one month of the end of the month in which the instant offense occurred; 26.8 percent had a subsequent offense in the same month or by the end of the first month or by the end of the second month.

[2]The post time period includes <u>all</u> status offenders, not just those who were eligible for the DSO project.

[3]The number of cases in the post time period drops as months "at risk" increase because all youths entering the court too late to have the full follow-up period (1 month, 2 months,...8 months) were excluded when calculating the recidivism rate for that particular follow-up period. Thus, for each of the months at risk, all youths included in that analysis had at least that many months of follow-up data.

proportion recidivating increases as the time "at risk" increases. This is because the percentage recidivating is cumulative. It includes those who had a subsequent offense at any time during the risk period, not just those recidivating within a particular month. Thus, the data for three months means that 33 percent of the pre-program status offenders had a subsequent offense within a followup period that extended for three months after the beginning of the month in which the instant offense occurred. It does not mean that 33 percent recidivated during the third month after the instant offense.

The difference between pre and post recidivism rates (column four of Table 14) increases from 2.7 percent in the same month to about 8 percent within three months and stabilizes at about 8 percent difference between pre and post as the risk period increases to eight months.

Although the differences observed would indicate that DSO had the effect of reducing recidivism, there are several other potential explanations of why recidivism was lower in the post time period. One possibility is that there was a downward trend in recidivism rates during the pre-program time period which simply continued after DSO began. Another alternative explanation is that the characteristics of status offenders were changing, over time or at the time that DSO began, and the difference in recidivism is attributable to the fact that the status offenders during the post time period did not have the same characteristics as status offenders during the pre-program phase.

The multiple regression analysis of pre and post data indicates that neither of these explanations accounts for the change in recidivism during the post time period. In Table 15 are the results of a multiple regression analysis using all pre and post cases that had at least three months of "at risk" time. The results show that the project intervention had a statistically

TABLE 15

MULTIPLE REGRESSION ANALYSIS OF DSO IMPACT ON 3-MONTH RECIDIVISM OF STATUS OFFENDERS, PRE AND POST[1]

No. of cases=2,285

INDEPENDENT VARIABLE[3]	DEPENDENT VARIABLE: RECIDIVISM WITHIN 3 MONTHS OF INSTANT OFFENSE[2] Zero Order Correlation	B	Beta	F Value	Probability
DSO Startup	-.14	-.08	-.08	6.3	P<.05
Monthly Trend	-.14	-.004	-.10	10.4	<.01
Number of Prior Status Offenses	.22	.07	.20	90	<.001
Number of Prior Delinquent Offenses	.16	.07	.14	42	<.001
Age (older)	-.08	-.03	-.11	28.5	<.001
Sex (female)	-.007	-.007	.00	.10	NS
R^2=.10					
F=30.7					

[1]The zero order correlation shows the relationship of each variable on the left with recidivism when no other variables are controlled. B is the unstandardized partial regression coefficient and beta is the standardized partial regression coefficient. The analysis was conducted on the juvenile court computerized data base.

[2]Cases which did not have at least a three month risk period were excluded. Otherwise, all status offenders in the post period, not just the DSO project youths, were included.

[3]DSO start-up is a dummy variable with pre-project cases having a score of zero and post-DSO status offenders a score of one. The interaction term (DSO times month) was not significant. Other characteristics of status offenders (family stability, school status) were not significant and were omitted from the equation.

significant effect in reducing recidivism, controlling for age, sex, number of status offense priors, number of delinquent priors, and the family stiuation of the youth. The change attributable to DSO was a shift in the level of recidivism rather than a shift in the trend. The trend, for the entire time period, was statistically significant but of very minor magnitude. Recidivism, on the average, declined by less than one-half of one percent per month. The average recidivism rate for the three-month followup, however, dropped by about seven percent when DSO began, even with the other variables held constant.

It has been shown previously in this report that the proportion of youths detained declined as a result of the DSO project and the proportion of status offenders on whom petitions were filed also dropped. A multiple regression analysis of the effect of petitions and detention on recidivism is shown in Table 16. The results indicate that youths who are detained are more inclined to recidivate than those who are not, even when prior offenses have been controlled along with age, sex, and so on. In contrast, youths on whom petitions are filed tend to recidivate at a lower rate than others, when priors and socio-economic characteristics have been controlled. (Somewhat different results are obtained in the post only analysis, however.) More important, as shown at the bottom of Table 16, the DSO intervention had a statistically significant impact on recidivism independent of its effect on detention and petitions.

The results of the multiple regression analysis are substantiated by an examination of recidivism (pre and post) for youths with different characteristics (Table 17). Regardless of whether a three or six month "at risk" time is used, the results show that recidivism rates within selected characteristics of the status offenders are uniformly lower during the post-program time period.

TABLE 16

EFFECT OF DETENTION & PETITIONS ON 3-MONTH RECIDIVISM

OF STATUS OFFENDERS, PRE & POST[1]

N=2,285

INDEPENDENT VARIABLE	DEPENDENT VARIABLE: RECIDIVISM WITHIN 2-MONTHS OF INSTANT OFFENSE				
	Zero Order Correlation	B	Beta	F Value	Probability
Petitions	-.02	-.05	-.05	5.6	<.05
Detention	.07	.06	.06	8.8	<.01
Number Prior Status Offenses	.22	.07	.20	87	<.001
Number Prior Delinquent Offenses	.16	.07	.12	33	<.001
Age	-.08	-.03	-.10	26	<.001
R^2=.07					
F=26					
DSO Intervention[2]	-.14	-.07	-.07	4.7	
Trend[2]	-.14	-.004	-.10	11	

[1]The zero order correlation shows the relationship of each variable on the left with recidivism when no other variables are controlled. B is the unstandardized partial regression coefficient and beta is the standardized partial regression coefficient. The analysis was conducted on the juvenile court computerized data base.

[2]The effect of DSO is estimated with all the other variables in the equation. The effect of petitions and detention (upper part of table) are estimated without the intervention variables being in the equation. Cases without at least a three-month risk period were excluded.

TABLE 17

THREE AND SIX MONTH RECIDIVISM RATES

OF STATUS OFFENDERS, PRE AND POST[1]

CHARACTERISTIC	Three Month Recidivism Rates		Six Month Recidivism Rates		NUMBER OF CASES			
					Three Months		Six Months	
	PRE	POST	PRE	POST	PRE	POST	PRE	POST
SEX								
Male	31	26.2	42	37	976	305	976	200
Female	32.1	24.6	42	33	1,354	422	1,354	263
LIVING SITUATION								
both natural parents	30.5	22.2	42	38	941	266	941	162
two parents, one step	36.7	21.9	46	32	327	137	327	79
one parent	31.0	29	41	25	591	209	591	134
other[2]	35.9	31	46	35	345	96	345	71
AGE								
12-13	32.5	19	45	26	379	108	379	66
14-15	36.7	30	46	41	1,147	380	1,147	242
16-17	24	22	35	29	738	219	738	146
OFFENSE								
Curfew	24.1	20	33	23	177	66	177	52
Runaway	32.5	24	40	35	1,093	329	1,093	205
Incorrigible	33	33	45	39	785	213	785	157
Truant	33	15	47	24	120	67	120	17

[CONTINUED ON NEXT PAGE]

TABLE 17 (continued)

CHARACTERISTIC	Three Month Recidivism Rates		Six Month Recidivism Rates		NUMBER OF CASES			
					Three Months		Six Months	
	PRE	POST	PRE	POST	PRE	POST	PRE	POST
NUMBER OF PRIOR STATUS OR DELINQUENT OFFENSES								
none	23	17	30	24	1,103	368	1,103	223
one	37	23	47	36	521	157	521	92
two	36	37	55	41	300	71	300	44
three	47	42	60	54	406	139	406	106
PETITIONS								
no petition filed	31	26	43	37	1,627	554	1,627	333
petition filed	33	23	40	30	703	175	703	132
DETAINED								
not detained	28	24	38	35	664	358	664	201
detained	33	27	44	35	1,666	371	1,666	264

[1]The analysis is based on Clark County computerized data, July 1976 through June 1977.

[2]"Other" includes relatives, group homes, foster homes, or institutions.

Comparison of Experimental and Control

Even though the previous analysis indicates that DSO had a significant impact on recidivism, it is important to ascertain whether the post-DSO change was due primarily to the experimental group or whether some (or all) of it could be attributed to the control and ineligible groups.

Table 18 contains data showing the proportion of youths within the experimental and control groups who recidivated within the same month as the instant offense, within one month of the instant offense, two months, and so on. The experimental group has lower recidivism rates for each of the different amounts of time "at risk." The differences become substantial enough after three months of followup (nine percent) to approach statistical significance at the .05 level and clearly are significant at or beyond that level when the risk period is four through eight months.

The differences observed in Table 18 could, of course, be due to different characteristics of the youths in the two groups because, as has been noted several times, the random assignment of youths to experimental and control groups was not perfectly adhered to and some differences exist between the two groups.

The data in Table 19 show the recidivism rates of experimental, control, and ineligible youths within each of several selected characteristics of the youths.

The recidivism rate within the experimental group for both the three-month and six-month followup periods is generally lower than that for the control group regardless of the age of the youth, the living situation, the type of offense, and the number of prior offenses (status or delinquent). For males within the experimental group the recidivism rate is slightly higher after three months at risk (25 percent compared to 21 percent within the control group), but is lower than the control group after six months

TABLE 18

COMPARISON OF EXPERIMENTAL & CONTROL GROUP RECIDIVISM RATES

FOR DIFFERENT LENGTHS OF FOLLOWUP TIME[1]

NUMBER MONTHS OF FOLLOWUP	Percent Recidivating (re-contact w/ court)		Z Value	Prob	Difference Between E & C	Number of Cases Included In Analysis	
	Exper	Contr				Exper	Contr
0	6.4%	10.2%	1.43	(.16)	3.8	362	127
1	15.5%	14.2%	.34	(.50)	1.3	330	127
2	18.5%	21.3%	.67	(.50)	2.8	297	127
3	20.1%	29.3%	1.82	(.06)	9.2	263	123
4	24.4%	37.9%	2.58	(.01)	13.5	217	116
5	26.3%	40.6%	2.44	(.01)	14.3	156	96
6	29.4%	48.0%	2.66	(.01)	18.6	126	75
7	33.0%	56.0%	2.9	(.01)	23.0	112	59
8	38.1%	57.0%	2.11	(.04)	18.9	84	49

[1]Recidivism is measured as a subsequent court contact for a delinquent or status offense after the instant status offense. Those who had no subsequent offenses within the risk period shown on the left are included as "non-recidivators" when calculating the percentage. The percentages are comulative across the risk period.

TABLE 19

COMPARISON OF EXPERIMENTAL, CONTROL, & INELIGIBLE RECIDIVISM RATES FOR THREE & SIX MONTHS OF TIME AT RISK[1]

CHARACTER-ISTICS	THREE MONTHS AT RISK			SIX MONTHS AT RISK			NUMBER OF CASES					
							THREE MONTHS			SIX MONTHS		
	Exper	Contr	Inelg	Exper	Contr	Inelg	Exper	Contr	Inelg	Exper	Contr	Inelg
AGE												
12-13	19	31	15	18	50	24	43	13	52	17	8	41
14-15	26	32	32	40	52	38	140	72	68	68	46	128
16-17	14	24	27	16	40	33	72	34	113	37	20	89
LIVING SITUATION												
both parents	20	29	21	40	45	33	109	48	109	52	31	79
two parents, one step	13	19	30	17	47	33	48	26	63	18	15	46
one parent	36	33	31	25	48	37	91	40	78	48	23	63
other[2]	23	43	32	(17)	(50)	36	13	7	76	6	4	61
OFFENSE												
curfew	(38)	(22)	16	(20)	(50)	22	8	9	49	5	2	45
runaway	16	28	28	23	46	36	113	50	166	44	35	126
incorrigible	29	36	36	32	51	40	91	50	72	60	37	60
truant	11	15	25	(33)	(0)	(14)	38	13	16	9	1	7
SEX												
male	25	21	28	34	50	36	112	29	164	50	14	136
female	17	32	27	26	48	31	149	94	179	74	61	128

[CONTINUED ON NEXT PAGE]

TABLE 19 (continued)

CHARACTER-ISTICS	THREE MONTHS AT RISK			SIX MONTHS AT RISK			NUMBER OF CASES					
							THREE MONTHS			SIX MONTHS		
	Exper	Contr	Inelg	Exper	Contr	Inelg	Exper	Contr	Inelg	Exper	Contr	Inelg
PRIOR OFFENSES												
none	19	24	13	26	42	17	151	70	147	69	41	113
one	22	40	19	39	64	24	73	25	53	36	14	42
two	32	24	48	25	40	50	25	17	29	12	10	22
three+	21	46	45	22	60	56	14	11	114	9	10	87
# STATUS OFFENSE PRIORS												
none	20	22	15	30	43	22	186	85	200	83	47	155
one	21	50	41	26	64	38	53	18	51	31	11	42
two	27	30	58	33	38	67	15	10	24	6	8	18
three+	33	50	43	33	67	55	9	10	68	6	9	49
# DELINQ. OFFENSE PRIORS												
none	20	31	22	28	47	24	213	102	210	103	64	160
one or more	24	19	36	35	55	49	50	21	133	23	11	104

[1]The analysis is based on Clark County computerized data, July 1976 through June 1977. Whether a youth was in the experimental, control, or ineligible group was determined from the data IPA collected for the USC national evaluation and this designation was added to the raw court data file.

[2]"Other" includes relatives, group homes, foster homes, or institutions.

at risk. (Tests of statistical significance have not been calculated for this table because its purpose is to examine whether the patterns of differences--9 percent lower for three months and almost 19 percent lower for six months--is maintained within various categories of youths.) In general, the evidence in Table 19 shows that the observed differences in Table 18 are not attributable to differences between the types of status offenders handled by the two groups.

This conclusion is further substantiated with the multiple regression analysis reported in Table 20. The treatment variable, even with all priors and socio-economic characteristics controlled, produces about a 10 percent reduction in the recidivism rate for a three-month "at risk" period and this is statistically significant (F=4.07) beyond the .01 level.

The effect of petitions and detention on recidivism, controlling for priors and socio-economic characteristics, is shown in Table 21, but the results (based only on a comparison of experimental and control group youths) differ from those found when the entire pre-post data were examined. For the former, it appears as if the filing of a petition increases the probability of recidivism, whereas detention is not significantly related to recidivism. For the entire pre-post data, detention had a significant relationship to higher recidivism, but petitions were related to lower recidivism. It should be noted that being in the experimental group (Table 21) maintains a significant relationship with lower recidivism even when detention and petitions are controlled.

A final question is whether some change in the community or at the court produced a change in the recidivism rates of all youths--status offenders and delinquents--and, therefore, the apparent effect of DSO has been confused with this outside influence on the system. An analysis of recidivism rates of delinquents shows 18 percent of youths whose instant offense

TABLE 20

MULTIPLE REGRESSION OF TREATMENT EFFECTS ON RECIDIVISM RATES FOR THREE MONTHS AT RISK TIME, EXPERIMENTAL VERSUS CONTROL[1]

N=345

	Zero Order Correlation	B	Beta	F Value	Probability
Treatment (experimental)	-.11	-.10	-.11	4.07	<.05
Prior status offenses	.10	.05	.09	3.2	<.05
Prior delinquent offenses	-.02	-.02	-.03	.28	NS
Parents	.08	.003	.07	1.78	NS
Age	-.05	-.01	-.05	.81	NS
Sex (female)	-.02	-.03	-.04	.61	NS
Constant		.46			R^2=.11

[1]The zero order correlation shows the relationship of each variable on the left with recidivism when no other variables are controlled. B is the unstandardized partial regression coefficient and beta is the standardized partial regression coefficient. The analysis was conducted on the juvenile court computerized data base.

TABLE 21

MULTIPLE REGRESSION OF PETITIONS & DETENTION WITH RECIDIVISM RATES FOR THREE MONTHS RISK TIME, EXPERIMENTAL & CONTROL GROUPS[1]

N=345

	Zero Order Correlation	B	Beta	F Value	Probability
Detention	.04	.04	.04	.48	NS
Petitions	.10	.17	.10	3.2	$<.05$
Prior Status Offenses	.11	.05	.09	2.69	$<.05$
Living Situation	.08	.003	.07	1.68	NS
Age	-.05	-.01	-.05	.70	NS
Prior Delinquent Offenses	-.02	-.03	-.03	.27	NS
Sex (female)	-.02	-.02	-.02	.174	NS
constant		.35			$R^2=.03$
Treatment[2] (control=1; experimental=0)	-.11	-.10	-.10	3.35	

[1]The zero order correlation shows the relationship of each variable on the left with recidivism when no other variables are controlled. B is the unstandardized partial regression coefficient and beta is the standardized partial regression coefficient. The analysis was conducted on the juvenile court computerized data base.

[2]The effect of the treatment is estimated with the other variables in the equation. In the upper part of the table, the effects are estimated without the treatment variable being controlled.

was a delinquency had a subsequent delinquent or status offense within three months during the pre-program time period compared with 19 percent of the post-DSO delinquents. When six months of followup are used, the results are quite similar: During the pre-program time period, delinquent offenses were followed by a subsequent status or delinquent offense in 22 percent of the cases compared with 24 percent recidivism for the post-DSO youths. Thus, the recidivism rates for delinquents did not change at all, or increased slightly, providing evidence that the observed decrease for status offenders was not produced by some outside factor influencing all youths in the community.

Discussion

The major conclusions from this section are:

1. The DSO intervention in July 1976 produced a statistically significant decrease in recidivism of status offenders.

2. The reduction in recidivism was due primarily to the experimental DSO youths who, when compared with the control group, had a significantly lower recidivism rate.

3. For a three-month followup period the pre-program recidivism rate of status offenders was 33 percent compared with 25 percent for the post-DSO status offenders (experimental, control, and ineligibles). A difference of about seven percent between pre and post recidivism rates was maintained even when a variety of possibly confounding variables were controlled (prior offenses, age, living situation, and sex). For a six-month followup period the differences between pre and post were 44 percent (pre) and 35 percent (post).

4. The experimental group recidivism rate for a three-month followup period was 20 percent compared with 29 percent for the control group. When

other possibly confounding variables were controlled the difference between the groups was about 10 percent. For a six-month followup the difference between experimental and control groups was much larger (29 percent compared with 48 percent).

5. The effect of detention and/or filing petitions on status offender recidivism is difficult to assess and disentangle from the effect of prior offenses. When the pre and post time periods are examined together, it appears as if recidivism increases if the youth is detained but declines if a petition is filed. For the experimental and control groups in the post time period, a different pattern was observed: Recidivism increased if a petition was filed but detentions had no effect. In either case, the effect was rather trivial (in the general area of 2 or 3 percent differences). The effect of DSO on recidivism was maintained even when both petitions and detention were statistically controlled in the regression equations.

SECTION 4.D.

Evaluation Report
TARGET HARDENING*

By

City of Seattle
Office of Policy Planning
Law and Justice Planning Office

Lawrence G. Gunn
Director

Molly Newcomb, Ph.D.
Kenneth E. Mathews, Jr., Ph.D.
JoAnne Pullen
Doris Lock

August, 1977

*This is the full text of the Target Hardening Evaluation Report.

INTRODUCTORY COMMENTS ON:

"TARGET HARDENING EVALUATION"

Many of the typical problems in assessing the effectiveness of target-hardening crime prevention efforts on burglary are discussed in the excerpts from this evaluation report. And, the techniques used to resolve them should be of interest to other evaluators.

One major problem in evaluations of this type is that the data used to measure burglaries suffers from problems of low reliability. And, validity of the official data is a problem if the concept being measured includes unreported burglaries as well as reported ones. The evaluator initially intended to use three sets of measures for burglary rates: Official police data, housing authority data from the target housing units, and results from two victimization surveys (one pre, one post). Several techniques were used to check the reliability and accuracy of the data and the evaluator concluded that the survey results contained too much error to use.

If the reliability of the survey data had not been checked carefully and if it had been used to assess project effectiveness, the conclusion would have been quite confusing and almost non-interpretable.

Another common problem in evaluation is the lack of rigorous experimental conditions. The technique used in the Target Hardening Evaluation was to utilize multiple indicators of performance; several different designs and comparison groups; and several different analysis techniques. If the findings from multiple tests of project effectiveness are generally consistent, the evaluator has far more confidence in the conclusions than would have been the case if project effectiveness had been assessed using only

one dependent variable, one design, and one analysis strategy.

A third problem in evaluating area-based crime prevention programs is that crimes may be displaced into other areas or one type of offense could be displaced to another (burglary to robberies, for example). These problems were anticipated before the evaluation began and data were collected to examine whether an unintended consequence of the project was to displace crimes.

The evaluation report also contains an interesting discussion of problems inherent in assessing change in rates when the initial rate is very low (floor effects).

Target Hardening Evaluation
Grant Award No. 1479
July 1, 1974, to February 29, 1976

The Target Hardening Project attempted to reduce burglary rates in four Seattle Housing Authority housing projects by making housing units more difficult for burglars to penetrate.

Specific hardening measures employed were:

1. installation of exterior solid-core doors or reinforcement of existing doors;

2. installation of one-inch dead-bolt locks on all exterior doors;

3. pinning of sliding glass windows to limit opening to less than nine inches;

4. construction of stub walls to prevent exterior access to interior door latches.

The goal of this project was to reduce, through target hardening, the incidence of burglaries committed in Seattle Housing Authority housing projects. This was to be achieved through deterrence by making forced entry physically more difficult and time-consuming, and, in cases of attempted or committed burglary, by leading to increased time for suspect observation.

Four specific objectives of the project were:

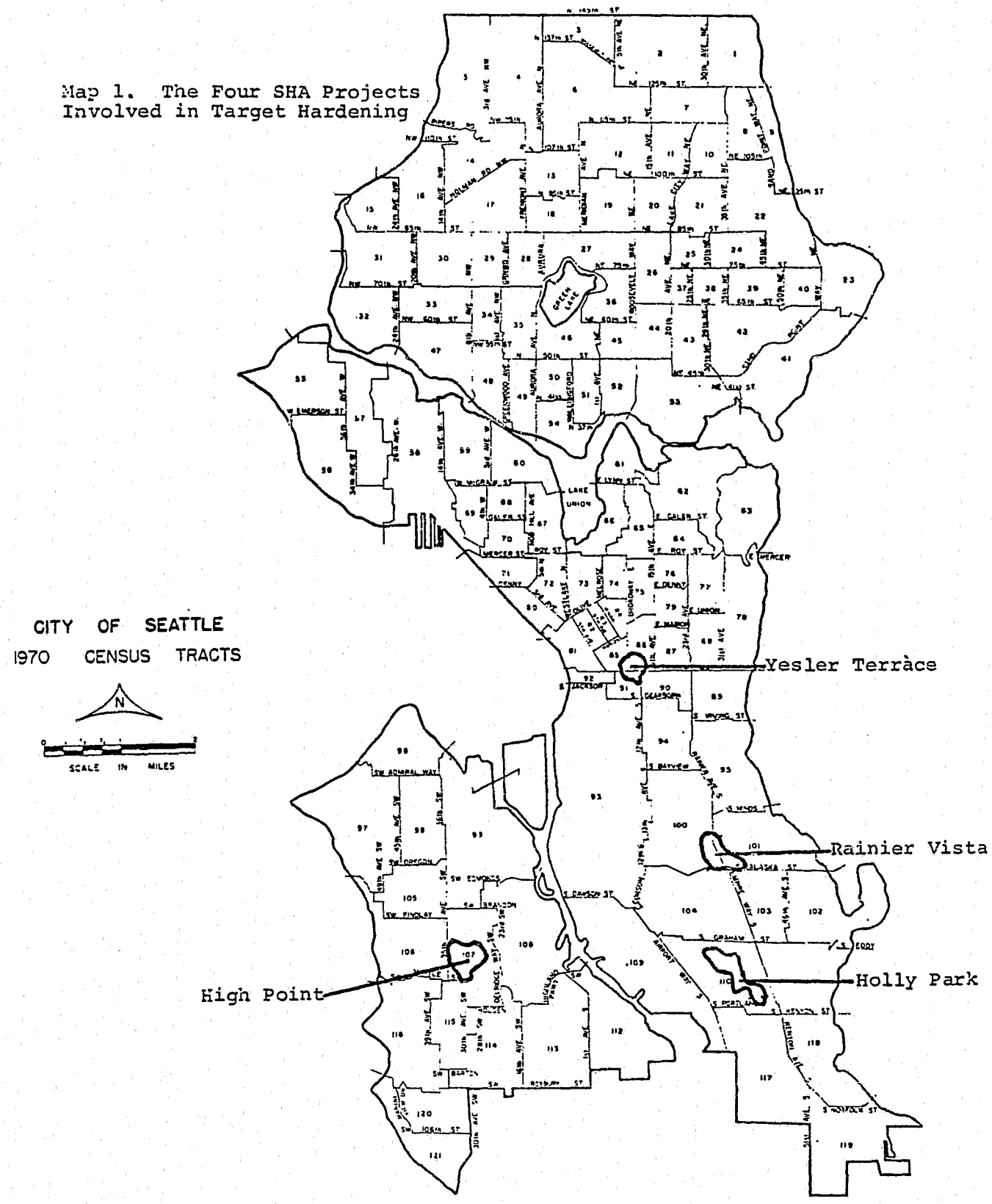
1. to effect significant reduction in the number of burglaries involving forced entry within the following Seattle Housing Authority housing projects: High Point, Holly Park, Rainier Vista, and Yesler Terrace;

2. to increase significantly the arrest-per-burglary rate within Seattle Housing Authority housing projects;

3. to increase significantly the proportion of witnessed burglaries involving forced entry into "hardened" housing units; (Rationale: The installation of solid core doors, one-inch dead-bolt locks, and the construction of walls or replacement with nonshattering material for all existing glass windows within a 32-inch radius of door latches will lead to increased noise and longer time periods required to make forced entry. This will result in increased exposure for offenders and a higher likelihood of being observed);

4. to increase significantly the proportion of witness-and/or victim-identified suspects of forced-entry burglaries into "hardened" housing units. (Rationale: Increased offender exposure time would enable witnesses to observe more and subsequently describe suspects in more detail.)

Map 1. The Four SHA Projects Involved in Target Hardening



The total Law Enforcement Assistance Administration cost for this project, including matching funds from the city and state, was $42,222. Of this amount, $35,111 was spent on materials and labor for the hardening measures described above. These LEAA funds were supplemented by Housing and Urban Development Modernization funds, amounting to $405,868.14.

In all, 3,082 living units were hardened. This total includes all permanent residential units in the High Point, Holly Park, Yesler Terrace, and Rainier Vista housing projects. The locations of these projects are shown on Map 1.

Hardening began in July, 1974, and was completed in May, 1975. For individual housing projects, hardening construction occurred as follows: High Point, December, 1974, through May, 1975; Holly Park, July, 1974, through May, 1975; Yesler Terrace, July, 1974, through November, 1974; Rainier Vista, July, 1974, through September, 1974.

For evaluation purposes, data on actual or attempted residential burglaries (hereafter referred to as "burglaries") were collected from three sources: reports to the Seattle Police Department (SPD data), reports to Seattle Housing Authority project managers (SHA data), and responses to crime victimization surveys conducted on random samples* of residents in late 1974 and again in late 1975 (victimization data). Also, data on robberies, thefts, and incidents of vandalism or arson were collected from SHA reports and victimization surveys.

To discover whether displacement of burglary to nearby areas would occur after hardening, SPD data on burglary and victimization data on burglary, robbery, theft, vandalism, and arson were collected for census tracts containing these four housing projects.** These data were subdivided according to type of housing within these census tracts: SHA project housing vs. non-SHA housing. Throughout this report, therefore, "non-SHA" housing refers to housing within the same census tracts as SHA housing but outside of the housing projects themselves. Crime rates in non-SHA housing in these census tracts and for Seattle as a whole provided comparison data for crime rates within the SHA projects being hardened.

Why use three different sources of data to find out burglary rates? Each data source has strengths and weaknesses. Seattle Police Department records show only those burglaries reported to police, not all which occur. SHA records exist only for SHA housing, and also require victims to take the initiative in reporting crimes. Residents vary in their tendencies to report to the SPD, to the SHA, to both, or to neither. Victimization data require the least effort on the part of the crime victim; he has only to answer the interviewer's questions. For this reason, victimization surveys usually show higher rates of crime than do statistics based on other crime reports. For example, a 1975 victimizzation survey of 13 American cities done by the Law Enforcement Assistance Administration found that only 52% of the burglaries had been reported to the police.***

---

* In 1974, residents of 228 SHA households and of 194 non-SHA households were interviewed. In 1975, residents of 303 SHA households and of 224 non-SHA households were interviewed.

** From Map 1 one can see that High Point Project lies in tract #107; Holly Park Project in tract #110; Yesler Terrace Project in tracts #85, 86, and 91; Rainier Vista Project in tracts #100 and 101.

*** Criminal Victimization Surveys in 13 American Cities, U.S. Department of Justice, LEAA, U.S. Government Printing Office, Washington, D.C., 1975.

A number of problems occurred with the victimization surveys for this evaluation, however. First, because target hardening took longer to complete than originally expected, the survey in 1975 provided data on only three months of post-hardening events. Such a short period of time can be unduly influenced by seasonal or chance events. Second, the victimization data were inconsistent with data from SPD and SHA sources, with victimization showing lower rates of burglary than had been reported to SPD and SHA in more than half of the comparisons of the rates. Third, the victimization survey asked respondents about crimes by which they had been victimized during the past year. Since SHA project residents are a mobile group, some respondents were reporting crimes which had occurred before they had moved to the project, while other potential respondents (who may have experienced crimes while in the project) had moved out before the survey. Fourth, the interviewers were residents of the SHA projects who did not have previous interviewing experience and showed some misunderstandings involving the purpose and content of the survey. Although a sample of interviews was verified, disclosing some falsified data, it was not financially possible to verify all interviews. A completely verified survey done by well-trained interviewers would probably have yielded more consistent results. Finally, the availability of data from two other sources made the victimization data supplementary but not essential for evaluation purposes.

Because of all the problems detailed above, the victimization data will not be presented in the main body of this report. For those interested, Appendix A provides a discussion of the data inconsistencies and summary statistics from these surveys.

Therefore, data from two sources, SPD and SHA, were used in evaluating the project's success in reaching the overall goal of burglary reduction and the first two of the four objectives outlined on page 1. Unfortunately, data regarding the last two objectives were not available from the SPD computer, so data could not be obtained without great effort and expense. Data from SPD and SHA sources were used to answer the following questions:

1. Did target hardening significantly reduce burglary rates in the four SHA housing projects? (This is related to Objective #1.)

2. Were there significant changes in burglary rates for non-SHA housing in these same census tracts for these time periods? (This question was asked to determine whether significant displacement of burglary from SHA housing to nearby non-SHA housing occurred as a result of target hardening within the SHA projects.)

3. How did burglary rate trends for SHA housing compare with trends for Seattle as a whole, and with trends for non-SHA housing?

4. How did the mode of entry used by burglars in entering SHA housing change after hardening was completed? (This question is relevant to Objective #3, but cannot be definitive for that objective.)

5. Did the proportion of burglaries cleared by police arrest change for SHA housing after hardening? How did the changes in SHA clearance rates compare with changes for Seattle as a whole and for non-SHA housing?

6. Did robbery and vandalism show any changes in rate during these time periods in SHA housing? (Was there displacement to other crimes within the housing projects?)

7. What was the overall result of target hardening?

To answer these questions, statistical tests were applied to differences in rates of occurrence before and after hardening. A difference or change was considered to be statistically significant if it showed a probability level below .05. This standard of significance is conventional for social science research; it implies that observed differences or changes could be due to chance occurrences less than five percent of the time.

Two time periods were used in comparing crime rates: pre-hardening vs. post-hardening time periods. The "pre-hardening" time period includes months prior to complete hardening of any single living unit; the "post-hardening" time period includes months following 67% completion of hardening (67% of the living units completely hardened.*) The number of months on which pre- and post-hardening averages are based differs by the source of the data. The number of months in each time period for which data were available is listed in Table 1.

DATA BASE FOR BURGLARY RATES BY DATA SOURCE AND TIME PERIOD

TABLE 1

	PRE-HARDENING	POST-HARDENING
	Months for which pre-hardening data were available	Months for which post-hardening data were available
SPD Data	January 1973 - June 1974 (18 months)	April 1975 - September 1976 (18 months)
SHA Data	July 1973 - June 1974 (12 months)	April 1975 - June 1976 (15 months)

* This definition of "post-hardening" was used because hardening of 67% of units was thought to have considerable impact on burglary; also, the use of 67% rather than 100% as a cut-off point provided a longer post-hardening period for data comparison.

Question 1: Did target hardening significantly reduce burglary rates in the four SHA housing projects?

Table 2 shows the rates of burglary per 100 households per year for SHA housing according to both SPD and SHA data. Reductions in burglary rates range from a decrease of 44.4% (SPD data) to 59.2% (SHA data).

The last column in Table 2 gives the mean or average difference in monthly burglaries across the compared time periods. This number was determined by pairing the same months in the two time periods (for example, April, 1974, with April, 1975) and subtracting the number of burglaries in the later month from the number of burglaries in the earlier month. (See Appendix B for an example of this process.) These differences were averaged to determine $\bar{D}$, the mean difference in monthly burglaries. If $\bar{D}$ is positive, that implies burglary rates have decreased; if $\bar{D}$ is negative, that implies burglary rates have increased during the time periods compared.

To see whether this mean change was significantly different from no change or not, an estimate of confidence limits for that specific $\bar{D}$ was made. The starred $\bar{D}$ numbers are different from zero with a less than 5 percent chance of error. For positive starred $\bar{D}$ numbers, this means that a significant decrease in burglary rates occurred.

According to SPD and SHA data, burglary rates in hardened SHA housing projects were significantly reduced from pre-hardening to post-hardening time periods.

In sum, burglary rates for the hardened SHA housing projects were significantly reduced after hardening was completed.

Table 3 provides a breakdown of these burglary rates by individual housing project, using both SPD and SHA data sources. The third column of this table shows decreases in burglary rates for all projects and data sources except for Yesler Terrace. Excluding Yesler Terrace, reductions in burglary rates range from 37.6 percent to 76.9 percent.

The fourth column shows the mean monthly differences scores and the results of confidence interval tests used to demonstrate the significance of these differences in relation to zero difference. These numbers were computed in the same way as the numbers in Table 2. Three of the eight mean difference scores are significantly greater than zero, indicating a significant decrease in burglary rate with a 5 percent level of chance error. Four more mean differences show decreases approaching significance, with a less than 10 percent level of chance error.

Why do SHA mean differences in High Point and Rainier Vista show significant reductions in burglary rates while SPD mean differences do not? Comparing SPD and SHA reported burglary rates, one sees that SHA rates are higher for seven of the eight projects and time periods. Only for the post-hardening time period at High Point does the SPD rate exceed the SHA rate. Thus it appears that a lower percentage of the burglaries are reported to the SPD than to the SHA. When reporting rates to the SPD are so low, it is difficult to measure change in burglary rates because the changes must affect those few people who will report

TABLE 2

AVERAGE RATES FOR BURGLARY - ALL SHA HOUSING

(Rates per 100 households per year)

	PRE	POST	(PRE TO POST) % CHANGE	MEAN DIFFERENCE
SPD Data	$\bar{X}$ = 5.88 S = 4.42 N = 18	$\bar{X}$ = 3.27 S = 1.27 N = 18	- 44.4%	$\bar{D}$ = 3.17* S_D = 4.86 $\bar{D}$ > 0 p < .05 N = 15
SHA Data	$\bar{X}$ = 11.13 S = 4.07 N = 12	$\bar{X}$ = 4.54 S = 1.48 N = 15	- 59.2%	$\bar{D}$ = 6.46* S_D = 4.09 $\bar{D}$ > 0 p < .05 N = 12

$\bar{X}$ = average rate per 100 households per year

N = number of months for which data were collected

S = standard deviation of the monthly averages

* significant at the .05 level

TABLE 3

AVERAGE RATES FOR BURGLARY BY HOUSING PROJECT - SHA HOUSING

(Rates per 100 households per year)

		PRE	POST	(PRE TO POST) % Change	MEAN DIFFERENCE
High Point	SPD Data	$\bar{X}$ = 7.15 S = 7.69 N = 24	$\bar{X}$ = 3.56 S = 2.09 N = 16	- 50.2%	$\bar{D}$ = 5.06 $S_{\bar{D}}$ = 9.37 $\bar{D}$ > 0 p < .10
	SHA Data	$\bar{X}$ = 10.42 S = 6.80 N = 17	$\bar{X}$ = 2.41 S = 1.10 N = 15	- 76.9%	$\bar{D}$ = 9.66* $S_{\bar{D}}$ = 8.05 $\bar{D}$ > 0 p < .05
Holly Park	SPD Data	$\bar{X}$ = 6.71 S = 3.96 N = 18	$\bar{X}$ = 4.19 S = 2.24 N = 16	- 37.6%	$\bar{D}$ = 3.05 $S_{\bar{D}}$ = 4.46 $\bar{D}$ > 0 p < .10
	SHA Data	$\bar{X}$ = 11.85 S = 7.91 N = 12	$\bar{X}$ = 4.33 S = 2.86 N = 15	- 59.2%	$\bar{D}$ = 7.15* $S_{\bar{D}}$ = 8.08 $\bar{D}$ > 0 p < .05
Yesler Terrace	SPD Data	$\bar{X}$ = 1.56 S = 2.12 N = 18	$\bar{X}$ = 1.90 S = 2.04 N = 21	+ 21.8%	$\bar{D}$ = -0.11 $S_{\bar{D}}$ = 3.10 $\bar{D}$ < 0 n.s.
	SHA Data	$\bar{X}$ = 5.33 S = 3.33 N = 12	$\bar{X}$ = 5.36 S = 5.28 N = 19	+ 0.6%	$\bar{D}$ = 2.50 $S_{\bar{D}}$ = 3.91 $\bar{D}$ > 0 p < .10
Rainier Vista	SPD Data	$\bar{X}$ = 4.90 S = 3.75 N = 18	$\bar{X}$ = 2.64 S = 2.80 N = 24	- 46.1%	$\bar{D}$ = 2.04 $S_{\bar{D}}$ = 4.04 $\bar{D}$ > 0 p < .10
	SHA Data	$\bar{X}$ = 14.69 S = 8.47 N = 12	$\bar{X}$ = 5.24 S = 3.48 N = 21	- 64.3%	$\bar{D}$ = 10.21* $S_{\bar{D}}$ = 9.50 $\bar{D}$ > 0 p < .05
TOTALS	SPD Data	$\bar{X}$ = 5.88 S = 4.42 N = 18	$\bar{X}$ = 3.27 S = 1.27 N = 18	- 44.4%	$\bar{D}$ = 3.17* $S_{\bar{D}}$ = 4.86 $\bar{D}$ > 0 p < .05
	SHA Data	$\bar{X}$ = 11.13 S = 4.07 N = 12	$\bar{X}$ = 4.54 S = 1.48 N = 15	- 59.2%	$\bar{D}$ = 6.46* $S_{\bar{D}}$ = 4.09 $\bar{D}$ > 0 p < .05

* significant at the .05 level

to SPD. Also, when reporting rates are lower, there is more chance for changes in reporting rates to affect data, and these reporting rate changes would not necessarily reflect true changes in actual burglary rates.

To sum it up, High Point, Holly Park, and Rainier Vista showed significant reductions in burglary rates after hardening was completed, as measured by one or both of SHA and SPD reports. Yesler Terrace showed no significant change on the basis of either data source.

Question 2: Were there significant changes in burglary rates for non-SHA housing in these same census tracts for these time periods?

Table 4 gives SPD burglary rates for all non-SHA housing in these census tracts for the same time periods used in the SHA housing statistics (see Table 1.) There are no SHA rates given because SHA does not compile statistics for non-SHA housing.

There is a small but significant decrease (8.9%) in the burglary rates for non-SHA housing from pre- to post-hardening. Thus burglary is not being displaced from the hardened SHA housing to nearby areas to any measurable degree. This conclusion is strengthened by comparing this 8.9% decrease in burglary rates for non-SHA housing with the city-wide trend in burglary rates for these time periods. City-wide, burglary rates were reduced by 5.8% during this time, so non-SHA areas had a somewhat greater reduction in burglary rates than did the city as a whole, providing more evidence that burglaries were not simply displaced to nearby areas by hardening.

TABLE 4

AVERAGE RATES FOR BURGLARY - ALL NON-SHA HOUSING

(Rates per 100 households per year)

	PRE	POST	(PRE TO POST) % Change	MEAN DIFFERENCE
SPD Data	$\bar{X}$ = 9.18 S = 1.22 N = 18	$\bar{X}$ = 8.36 S = 1.10 N = 18	- 8.9%	$\bar{D}$ 1.08* S_D = 1.34 $\bar{D}$ > 0 p < .05 N = 15

Table 5 shows the burglary rates in non-SHA housing, subdivided by project area. Housing near High Point and Holly Park shows a reduction in burglary rates; housing near Yesler Terrace and Rainier Vista shows an increase. The changes in non-SHA housing burglary rates for the areas surrounding Holly Park (a decrease) and Rainier Vista (an increase) are significantly different from no change, using a .05 level of confidence.

TABLE 5

AVERAGE RATES FOR BURGLARY IN NON-SHA HOUSING, SEPARATELY BY PROJECT AREA
(Rates per 100 households per year)

		PRE	POST	(PRE TO POST) % Change	MEAN DIFFERENCE
High Point Census Tract non-SHA housing	SPD Data	$\bar{X}$ = 19.24 S = 11.35 N = 24	$\bar{X}$ = 12.59 S = 7.55 N = 16	- 34.6%	$\bar{D}$ = 7.12 $S_{\bar{D}}$ = 14.04 $\bar{D} > 0$ p < .10
Holly Park Census Tract non-SHA housing	SPD Data	$\bar{X}$ = 30.58 S = 5.95 N = 18	$\bar{X}$ = 19.60 S = 5.44 N = 16	- 35.9%	$\bar{D}$ = 12.59* $S_{\bar{D}}$ = 10.36 $\bar{D} > 0$ p < .05
Yesler Terrace Census Tracts non-SHA housing	SPD Data	$\bar{X}$ = 5.50 S = 1.11 N = 18	$\bar{X}$ = 5.78 S = 1.62 N = 21	+ 5.1%	$\bar{D}$ = -0.32 $S_{\bar{D}}$ = 2.21 $\bar{D} < 0$ n.s.
Rainier Vista Census Tracts non-SHA housing	SPD Data	$\bar{X}$ = 6.70 S = 1.43 N = 18	$\bar{X}$ = 8.63 S = 1.86 N = 24	+ 28.8%	$\bar{D}$ = -1.75* $S_{\bar{D}}$ = 2.52 $\bar{D} < 0$ p < .05
TOTALS non-SHA housing	SPD Data	$\bar{X}$ = 9.18 S = 1.22 N = 18	$\bar{X}$ = 8.36 S = 1.10 N = 18	- 8.9%	$\bar{D}$ = 1.08* $S_{\bar{D}}$ = 1.34 $\bar{D} > 0$ p < .05

* significant at the .05 level

In summary, non-SHA housing shows a somewhat mixed picture of burglary rate changes, with the overall trend showing a small reduction in rates. While SPD burglary rates show a significant decrease in burglaries in the post hardening period for total non-SHA housing and the area around Holly Park shows a significant decrease in burglary rates, the Rainier Vista area shows a significant increase in burglaries, while the High Point and Yesler Terrace areas show no significant change.

Question 3: How did burglary rate trends for SHA housing compare with trends for Seattle as a whole and with trends for non-SHA housing?

Table 6 shows the averages, percentage changes, and mean differences for pre- and post-hardening burglary rates in SHA housing, non-SHA housing, and for Seattle as a whole.

Hardened SHA housing shows large and significant reductions in burglary rates after hardening, ranging from a 44.4% reduction (SPD data) to a 59.2% reduction (SHA data). Non-SHA housing shows a smaller but still significant reduction of 8.9%, while the city-wide reduction of 5.8% did not represent a significant change.

In conclusion, burglary rates for SHA housing showed more favorable trends than did rates for non-SHA housing or for Seattle as a whole during these time periods.

TABLE 6

COMPARISON OF BURGLARY RATES FOR SHA HOUSING, NON-SHA HOUSING, AND ALL OF SEATTLE

	AVERAGE RATES PER 100 HOUSEHOLDS PER YEAR		(PRE TO POST) % CHANGE		MEAN DIFFERENCE	
	SHA	non-SHA	SHA	non-SHA	SHA	non-SHA
SPD Data	Pre: 5.88 Post: 3.27	Pre: 9.18 Post: 8.36	-44.4%	- 8.9%	$\bar{D}$ = 3.17* S_D = 4.86 $\bar{D}$ > 0 p < .05 N = 15	$\bar{D}$ = 1.08* S_D = 1.34 $\bar{D}$ > 0 p < .05 N = 15
SHA Data	Pre: 11.13 Post: 4.54		-59.2%		$\bar{D}$ = 6.46* S_D = 4.09 $\bar{D}$ > 0 p < .05 N = 12	

	Average Rates	% Change	Mean Difference
All Seattle (SPD Data)	Pre: 4.84 Post: 4.56	- 5.8%	$\bar{D}$ = 0.28 S_D = 0.61 $\bar{D}$ > 0 n.s. N = 12

* significant at the .05 level

Table 7 shows the significance of these differences between burglary rate trends in SHA housing as compared with trends for Seattle and as compared with trends for non-SHA housing. These differences in trends were compared by performing a paired t-test on monthly differences in rates for SHA housing as compared with Seattle, and for SHA housing as compared with non-SHA housing; (see Appendix C for an example of these calculations.)

Table 7 shows that SHA housing had a significantly greater decrease in burglary rates than did Seattle as a whole. However, the decrease for SHA housing was not significantly greater than the decrease for non-SHA housing, as measured by SPD data. This SHA vs. non-SHA comparison approached significance, however, reaching a significance level of .10.

To sum it up, SHA housing showed a significantly greater reduction in burglary as compared with total Seattle rates, while the comparison between SHA and non-SHA housing approached significance.

TABLE 7

SHA vs. Seattle Burglary Rate Changes;
SHA vs. non-SHA Burglary Rate Changes;
Statistical Comparisons of Hardening Effects

	SHA Mean Difference	SEATTLE Mean Difference	Which housing had more favorable % change?	Was this SHA vs. Seattle difference significant?
SHA Housing vs. SEATTLE	$\bar{D}$ = 3.17* S_D = 4.86 $\bar{D} > 0$ $p < .05$ N = 15	$\bar{D}$ = 0.28 S_D = 0.61 $\bar{D} > 0$ n.s. N = 12	SHA	YES $\bar{D}$ = 3.48* S_D = 5.04 $p < .05$ N = 12

	SHA Mean Difference	NON-SHA Mean Difference	Which housing had more favorable % change?	Was this SHA vs. NON-SHA difference significant?
SHA Housing vs. Non-SHA Housing	$\bar{D}$ = 3.17* S_D = 4.86 $\bar{D} > 0$ $p < .05$ N = 15	$\bar{D}$ = 1.08* S_D = 1.34 $\bar{D} > 0$ $p < .05$ N = 15	SHA	NO $\bar{D}$ = 2.09 S_D = 4.52 $p < .10$ N = 15

* significant at the .05 level

Burglary rate changes for SHA and non-SHA housing, subdivided by project area, are given in Table 8. For comparison, figures for Seattle as a whole are provided at the bottom of Table 8. In the High Point, Holly Park, and Rainier Vista areas, SHA housing showed greater percentage decreases in burglary rates than did non-SHA housing. For the Yesler Terrace area, SHA housing showed a greater increase in burglary rate than did non-SHA housing. The reported burglary rates for the Yesler Terrace project are only one-third the lowest reported rates for any other area, however, leading one to suspect that under-reporting of burglary in this project may make these rates undependable.

Thus three of the four housing projects showed greater percentage decreases in burglary rates than did the corresponding surrounding non-SHA housing in each of these three areas.

In Table 8, the mean differences for non-SHA housing are greater than the corresponding differences for SHA housing, except in the case of Rainier Vista where the mean differences are approximately equal for both types of housing. This is true in spite of the fact that for each of the four areas, SHA housing shows greater percentage change than does non-SHA housing. The absolute value of the mean difference appears to be positively correlated with the value of the pre-hardening burglary rate; that is, initially higher burglary rates are likely to be changed by a larger amount than are initially lower burglary rates. A correlation between pre-hardening burglary rates and mean differences (all mean differences treated as positive) for the 12 measures in Table 8 yields a correlation coefficient of +.80, significant at the .001 level (for 10 degrees of freedom). This means that there is a highly significant correlation between the size of the pre-hardening burglary rate and the size of the mean difference found for that set of data.

Why should higher burglary rates pre-hardening be related to greater changes in burglary rates? Possible explanations include differential reporting rates and floor effects. When a smaller percentage of burglaries is reported, it is harder to document actual change because the change in burglary rates may not affect the small group of people who do the reporting. Floor effects upon change occur when a rate of occurrence is so low that increased effort is needed to lower it further. For example, it is easier to reduce the percentage of people lacking swine flu immunity when 55 percent of people lack immunity than when 5 percent of the people lack immunity. Similarly, if the initial burglary rate is 1.56 per 100 households per year (as reported in Yesler Terrace SPD data), it would be impossible to reduce this rate by 2 burglaries per 100 households per year unless you invent negative burglaries.

Support for both of the above hypotheses can be found in Table 8. SHA project residents appear to under-report burglaries to the SPD, because pre-hardening burglary rates from SHA data are higher than rates from SPD data for each project. Also according to SPD data, SHA burglary rates are markedly lower than such rates in immediately surrounding non-SHA housing for each project, also indicating under-reporting. While under-reporting apparently occurs for each housing project, floor effects seem to be involved as well because the correlation between pre-hardening burglary rates and mean differences remains significant (+.74, $p < .05$, $df=6$) when non-SHA housing is excluded from this correlation.

TABLE 8 COMPARISON OF SHA AND NON-SHA BURGLARY RATES BY PROJECT AREA

	Rates per 100 households per year		(Pre to Post) % Change		Mean Difference	
	SHA	non-SHA	SHA	non-SHA	SHA	non-SHA
HIGH POINT AREA SPD Data	Pre: 7.15 Post: 3.56	Pre: 19.24 Post: 12.59	-50.2%	-34.6%	$\bar{D}$ = 5.06 S_D = 9.37 $\bar{D} > 0$ $p < .10$	$\bar{D}$ = 7.12 S_D = 14.04 $\bar{D} > 0$ $p < .10$
SHA Data	Pre: 10.42 Post: 2.41		-76.9%		$\bar{D}$ = 9.66* S_D = 8.05 $\bar{D} > 0$ $p < .05$	
HOLLY PARK AREA SPD Data	Pre: 6.71 Post: 4.19	Pre: 30.58 Post: 19.60	-37.6%	-35.9%	$\bar{D}$ = 3.05 S_D = 4.46 $\bar{D} > 0$ $p < .10$	$\bar{D}$ = 12.59* S_D = 10.36 $\bar{D} > 0$ $p < .05$
SHA Data	Pre: 11.85 Post: 4.83		-59.2%		$\bar{D}$ = 7.15* S_D = 8.08 $\bar{D} > 0$ $p < .05$	
YESLER TERRACE AREA SPD Data	Pre: 1.56 Post: 1.90	Pre: 5.50 Post: 5.78	+21.8%	+ 5.1%	$\bar{D}$ =-0.11 S_D = 3.10 $\bar{D} < 0$ n.s.	$\bar{D}$ =-0.32 S_D = 2.21 $\bar{D} < 0$ n.s.
SHA Data	Pre: 5.33 Post: 5.55		+4.1%		$\bar{D}$ = 2.50 S_D = 3.91 $\bar{D} > 0$ $p < .10$	

* significant at the .05 level

TABLE 8 (continued)

Comparison of SHA and non-SHA Burglary Rates by Project Area

	Rates per 100 households per year		(Pre to Post) % Change		Mean Difference	
	SHA	non-SHA	SHA	non-SHA	SHA	non-SHA
RAINIER VISTA AREA SPD Data	Pre: 4.90 Post: 2.64	Pre: 6.70 Post: 8.63	-46.1%	+28.8%	$\bar{D} = 2.04$ $S_D = 4.04$ $\bar{D} > 0$ $p < .10$	$\bar{D} = -1.75^*$ $S_D = 2.52$ $\bar{D} < 0$ $p < .05$
SHA Data	Pre: 14.69 Post: 5.24		-64.3%		$\bar{D} = 10.21^*$ $S_D = 9.50$ $\bar{D} > 0$ $p < .05$	

	Rates per 100 households per year SEATTLE	(Pre to Post) % Change SEATTLE	Mean Difference SEATTLE
All Seattle SPD Data	Pre: 4.84 Post: 4.56	- 5.8%	$\bar{D} = 0.28$ $S_D = 0.61$ $\bar{D} > 0$ n.s. N = 12

* significant at the .05 level

Table 9 shows the significance of the differences between SHA and non-SHA housing burglary rate trends. This was measured by a paired t-test of differences, as was done in Table 7 above.

For High Point SHA and non-SHA areas, for example, Table 9 shows a mean decrease in burglary rates of 5.06 for SHA housing and of 7.61 for non-SHA housing. In both cases, these rate decreases are based on 16 pre-hardening months matched with 16 post-hardening months by month name, so N = 16. Although SHA housing showed a greater percentage decrease in burglary rates, there was no significant difference found between SHA and non-SHA rate changes. Thus, decreases in burglary in the High Point project could be a reflection of an area trend towards fewer burglaries, as well as an effect of target hardening.

In the Holly Park area, SHA housing and non-SHA housing both showed significant decreases in burglary rates, with percentage reductions nearly equal but slightly favoring SHA housing. The average decrease in burglary rate for SHA housing was 3.05; for non-SHA housing, it was 12.59*. A test for difference between these trends showed the non-SHA housing to have a significantly greater reduction than the SHA housing. Thus in the Holly Park area, SHA housing showed less of a decrease than did non-SHA housing.

In the Yesler Terrace area, there were no significant changes in the burglary rates for either SHA or non-SHA housing, and also no significant difference between the rate changes for the two types of housing.

For the Rainier Vista area, SHA housing showed a nearly significant decrease in burglary rates, while non-SHA housing showed a significant <u>increase</u>. The difference between these trends was significant, showing that the Rainier Vista project was not following the upward trend for burglary rates in non-SHA housing in the census tracts in which the project is located.

To sum it up, the High Point and Yesler Terrace areas showed no difference in burglary rate trends between SHA and non-SHA housing. The Holly Park and Rainier Vista areas showed significant differences in burglary rate trends for SHA versus non-SHA housing. For Holly Park, non-SHA housing had a greater decrease in burglary rates than did SHA housing; for Rainier Vista, non-SHA housing had a greater <u>increase</u> in burglary than did SHA housing in that area.

---

* The percentage reduction is greater for SHA housing even though the absolute change in numbers is less for SHA housing. The reason for this seeming contradiction is that the initial burglary rates for non-SHA housing in the Holly Park area is <u>much</u> higher than for SHA housing in this area.

TABLE 9

SHA VS. NON-SHA BURGLARY RATE CHANGES:
Statistical Comparison of Pre- vs. Post-Hardening and SHA vs. non-SHA, subdivided by project area

	Mean Difference SHA	Mean Difference non-SHA	Which housing had more favorable % change?	Was this SHA vs. non-SHA difference significant?
High Point Area (SPD Data)	$\bar{D}$ = 5.06 S_D = 9.37 $\bar{D} > 0$ $p < .10$	$\bar{D}$ = 7.61 S_D = 14.86 $\bar{D} > 0$ $p < .10$	SHA	NO $\bar{D}$ = -2.55 S_D = 10.30 n.s.
Holly Park Area (SPD Data)	$\bar{D}$ = 3.05 S_D = 4.46 $\bar{D} > 0$ $p < .10$	$\bar{D}$ = 12.59* S_D = 10.36 $\bar{D} > 0$ $p < .05$	SHA	YES $\bar{D}$ = -9.54* S_D = 11.06 $p < .05$
Yesler Terrace (SPD Data)	$\bar{D}$ = -0.11 S_D = 3.10 $\bar{D} < 0$ n.s.	$\bar{D}$ = -0.32 S_D = 2.21 $\bar{D} < 0$ n.s.	non-SHA	NO $\bar{D}$ = 0.21 S_D = 4.02 n.s.
Rainier Vista (SPD Data)	$\bar{D}$ = 2.04 S_D = 4.04 $\bar{D} > 0$ $p < .10$	$\bar{D}$ = -1.75* S_D = 2.52 $\bar{D} < 0$ $p < .05$	SHA	YES $\bar{D}$ = 3.79* S_D = 4.28 $p < .05$

* significant at the .05 level

Question 6: Did robbery and vandalism show any changes in rate during these time periods in SHA housing? (Was there displacement to these other crimes within the housing projects?)

Reports of robbery and vandalism during the pre- and post-hardening time periods were obtained from SHA housing managers. The reported rates for these crimes are given in Table 13.

Robbery was selected because, like burglary, it is a crime of economic gain. Thus it is reasonable to assume that a criminal unable to enter hardened housing units might turn to robbery to get the valuables he wants. Theft is even closer to burglary because it combines an economic motive with a lack of personal violence. Unfortunately, statistics on theft were not available. Vandalism was considered because attempted forced entry may be reported as vandalism.

Table 13 shows that the reported rates for both robbery and vandalism went down after hardening, showing that no displacement from burglary to these other crimes was apparent. In the case of robbery, a t-test on pre- and post-hardening average rates showed the decrease to be significant, while the decrease in vandalism was not significant.

In summary, there were no indications that hardening had resulted in a displacement to other crimes within the housing projects, as both robbery and vandalism within the projects decreased after hardening.

TABLE 13

AVERAGE RATES OF ROBBERY AND VANDALISM FOR SHA HOUSING
(Rates per 100 households per year)

	PRE	POST	t-SCORE
ROBBERY (SHA Data)	$\bar{X}$ = 1.78 S = 1.16 N = 12	$\bar{X}$ = 0.23 S = 0.21 N = 5	t = 2.80* p < .05 df = 15
VANDALISM (SHA Data)	$\bar{X}$ = 12.33 S = 12.48 N = 12	$\bar{X}$ = 0.86 S = 0.51 N = 5	t = 1.96 n.s. df = 15

* significant at the .05 level

Question 7: What was the overall result of target hardening?

Target hardening produced a significant reduction in burglary rates for hardened SHA housing. This reduction ranged from 44.4% (SPD reports) to 59.2% (SHA reports). This reduction in burglary compares favorably with a 5.8% reduction for Seattle as a whole during these time periods.

Hardening of the projects did not displace burglary into surrounding areas. Non-SHA housing in the same census tracts as the hardened SHA projects showed an 8.9% decrease in burglary rates after hardening. This decrease exceeds the 5.8% decrease for Seattle as a whole.

Specifically, High Point, Holly Park, and Rainier Vista projects all showed significant reductions in burglary rates from pre- to post-hardening, according to SPD or SHA data sources. Yesler Terrace, which had an extremely low pre-hardening burglary rate, showed no significant change in rate after hardening, using the same data sources.

Burglar's mode of entry for these census tracts was tabulated for pre- and post-hardening time periods, using SPD reports. Hardening was successful in decreasing the percentage of burglaries by forcible means: after hardening, a significantly higher percentage of burglaries were perpetrated through unlocked doors or windows than before hardening. Significant reductions in entries through doors reflected hardening's emphasis on door security, while the increase in entries by means of removing glass or frame slows down the entry process, making burglars more conspicuous and observable.

Clearance rates for burglaries in hardened SHA housing increased slightly but not significantly.

There were decreases in robbery and vandalism rates for hardened SHA housing, indicating that no displacement from burglary to these crimes took place after hardening.

APPENDIX A: VICTIMIZATION SURVEY DATA RESULTS AND PROBLEMS

As mentioned in the body of this report, data from the victimization surveys presented a number of problems: some interviews were falsified, interviewers were inexperienced, only a sample of interviews could be verified, and the later survey included only three months of data after hardening was 67% completed. In addition, the data from these surveys were inconsistent with reports from SHA and SPD data sources.

Table A-1 shows the contradictions between the trends shown by victimization survey data and SPD and SHA data. For this comparison, the SPD and SHA rates were based on the same months used in the victimization survey data. The pre-hardening months were July, 1973 through June, 1974; the post-hardening months were March, 1975 through June, 1975. The victimization data show considerable increases in burglary for both SHA and non-SHA housing; SPD and SHA data show considerable decreases in burglary for SHA and non-SHA housing over the same time periods. For the victimization data to be right and the other two data sources to be wrong, one would have to assume that some other irrelevant factor had affected reporting to both SPD and SHA simultaneously.

As mentioned on page 3, victimization surveys usually show higher rates of burglary than do data sources which depend upon victims' reports (such as do SPD and SHA). The pre-hardening victimization data display the opposite relationship with SPD and SHA data, as shown by the rates in Table A-1. For example, the pre-hardening burglary rate for SHA housing as shown by victimization data is 6.58, but the SPD rate is 7.07 and the SHA rate is 11.13. For non-SHA housing, the pre-hardening victimization data rate is 5.13, but the SPD data rate is 9.52. In all of these comparisons of pre-hardening rates, the victimization survey showed lower burglary rates than did the other data sources.

For post-hardening rates, the expected relationship between victimization and other data sources appears: victimization data rates are uniformly higher than are SPD and SHA rates.

Therefore, it seems that considerable under-reporting or unreliable data occurred in the earlier (pre-hardening) victimization survey. The later survey shows the expected relationships with other data sources, but the artifically low rate of the earlier survey creates an "increase" in burglary rates when both surveys are compared.

TABLE A-1

SHA VS. NON-SHA BURGLARY RATE CHANGES

Victimization Survey Data:

	PRE	POST	% CHANGE	MEAN DIFFERENCE
SHA Housing (Victimization Data)	$\bar{X}$ = 6.58 S = 5.16 N = 12	$\bar{X}$ = 7.92 S = 10.46 N = 3	+ 20.4%	$\bar{D}$ = 0.85 $S_{\bar{D}}$ = 8.79 $\bar{D}$ > 0 n.s. N = 3
NON-SHA HOUSING (Victimization Data)	$\bar{X}$ = 5.13 S = 5.13 N = 12	$\bar{X}$ = 10.71 S = 5.36 N = 3	+108.8%	$\bar{D}$ = -4.53 $S_{\bar{D}}$ = 10.00 $\bar{D}$ < 0 n.s. N = 3

SPD AND SHA DATA FOR SAME TIME PERIODS

	PRE	POST	% CHANGE	MEAN DIFFERENCE
SHA Housing (SPD Data)	$\bar{X}$ = 7.07 S = 4.98 N = 12	$\bar{X}$ = 3.50 S = 0.78 N = 3	- 50.8%	$\bar{D}$ = 2.00 $S_{\bar{D}}$ = 5.29 $\bar{D}$ > 0 n.s. N = 3
SHA Housing (SHA Data)	$\bar{X}$ = 11.13 S = 4.07 N = 12	$\bar{X}$ = 4.28 S = 1.35 N = 3	- 61.3%	$\bar{D}$ = 6.33 $S_{\bar{D}}$ = 4.73 $\bar{D}$ > 0 n.s. N = 3
NON-SHA HOUSING (SPD Data)	$\bar{X}$ = 9.52 S = 1.28 N = 12	$\bar{X}$ = 7.64 S = 0.40 N = 3	- 19.8%	$\bar{D}$ = 11.00 $S_{\bar{D}}$ = 5.29 $\bar{D}$ > 0 n.s. N = 3

Table A-2 shows the victimization data for rates of robbery, theft, and vandalism/arson for SHA and non-SHA housing. For crimes which are as infrequent as robbery, such small samples are not good indications of the frequency of the crime in the total population. For such a sample, one crime may make a tremendous difference in average rates. Unfortunately, there are no comparative data for non-SHA housing, as SPD data were not obtained for these crimes. For SHA housing, SHA reports indicate a decrease in robbery and a decrease in vandalism during these time periods, (see page 23 of this report), while the victimization data indicate no change for these crimes in SHA housing. If one can assume under-reporting for these crimes similar to under-reporting of burglary during the pre-hardening victimization survey, the pre-hardening figures in Table A-2 are underestimates. Therefore, there may have been a decrease masked by the under-reporting problem of the earlier survey. Similarly, the under-reporting during the earlier survey may have caused the apparent increase in theft post-hardening.

TABLE A-2

MEAN RATES OF ROBBERY, THEFT, AND VANDALISM/ARSON
Victimization Survey Data

	PRE	POST	t-SCORE
ROBBERY			
SHA Housing	$\bar{X}$ = 0.00 S = 0.00 N = 12	$\bar{X}$ = 0.00 S = 0.00 N = 3	t = 0.00 n.s. df = 13
NON-SHA Housing	$\bar{X}$ = 1.55 S = 2.78 N = 12	$\bar{X}$ = 0.00 S = 0.00 N = 3	t = 0.73 n.s. df = 13
THEFT			
SHA Housing	$\bar{X}$ = 1.32 S = 2.37 N = 12	$\bar{X}$ = 3.96 S = 3.96 N = 3	t = 1.38 n.s. df = 13
NON-SHA Housing	$\bar{X}$ = 0.49 S = 1.79 N = 12	$\bar{X}$ = 7.12 S = 8.20 N = 3	t = 2.39* p < .05 df = 13
VANDALISM/ARSON			
SHA Housing	$\bar{X}$ = 2.64 S = 2.76 N = 12	$\bar{X}$ = 2.64 S = 4.56 N = 3	t = 0.00 n.s. df = 13
NON-SHA Housing	$\bar{X}$ = 2.04 S = 3.00 N = 12	$\bar{X}$ = 1.80 S = 3.12 N = 3	t = 0.12 n.s. df = 13

* significant at the .05 level

Table A-3 gives the victimization survey data on burglary reporting rates. Each respondent in the 1974 and 1975 victimization surveys who reported being victimized by one or more crimes was asked whether each crime had been reported to the police. Respondents living in SHA housing were also asked whether each crime had been reported to the manager of the project.

TABLE A-3

BURGLARY REPORTING RATES TO SPD AND SHA

Victimization Survey Data

TIME PERIOD	% BURGLARIES REPORTED		
	To SHA by SHA residents	To SPD by SHA residents	To SPD by non-SHA residents
PRE (12 months)	63.2% (12 of 19)	61.1% (11 of 18)	81.8% (9 of 11)
POST (3 months)	70.0% (14 of 20)	76.2% (16 of 21)	71.8% (28 of 39)

Table A-4 shows the inconsistencies in reporting rates as found from victimization data and from actual reports of burglaries to SPD and SHA. According to victimization data, pre-hardening non-SHA burglary rates are 5.13 (see Table A-4) and 81.8% of these burglaries were reported to SPD (Table A-3). Thus the survey-estimated rate for SPD-reported burglary is 5.13 times .818, or 4.20. However, the actual reports of burglaries to SPD gave an SPD rate of 9.52, signifying that considerably more burglaries are reported to SPD than victimization survey data would imply. This means respondents are claiming to have reported fewer burglaries than were actually reported. This is the opposite of what one would expect. The pre-hardening victimization survey again shows a marked under-reporting of burglary data; the survey-estimated rates are markedly and consistently below the actual SPD and SHA rates for the pre-hardening time period as shown in Table A-4.

For the post-hardening victimization survey, the data show a much more reasonable relationship: respondents claim to have reported more burglaries to SPD and SHA than were actually reported, and the actual SPD and SHA rates do not differ as markedly from the survey estimate of these rates as for the pre-hardening victimization survey data.

To sum it up, the pre-hardening victimization survey shows contradictions and inconsistencies indicating a lack of reliability and validity in the data, possibly due to serious under-reporting of burglary. The post-hardening victimization survey data show more reasonable relationships with data from other sources. Unfortunately one needs reasonable data from both time periods to indicate change due to hardening.

TABLE A-4

VICTIMIZATION DATA: REPORTING RATE INCONSISTENCIES FOR BURGLARY DATA

SPD Data and Victimization Data

TIME PERIOD	SHA HOUSING			NON SHA HOUSING		
	Victimization Data	Actual SPD Data	Survey-Estimated SPD Data	Victimization Data	Actual SPD Data	Survey-Estimated SPD Data
PRE (12 months)	6.58	7.07	4.02	5.13	9.52	4.20
POST (3 months)	7.92	3.50	6.04	10.71	7.64	7.69

SHA DATA (for SHA Housing Only) and VICTIMIZATION DATA

TIME PERIOD	Victimization Data	Actual SHA Data	Survey-Estimated SHA Data
PRE (12 months)	6.58	11.13	4.16
POST (3 months)	7.92	4.28	5.54

Victimization Survey Time Periods Used for all data sources

SECTION 4.E.

Seattle Community Accountability Program (formerly Youth Service Bureau System) CRIME IMPACT AND 12-MONTH RECIDIVISM ANALYSIS*

By

Seattle Law and Justice Planning Office

Kenneth E. Mathews, Jr., Ph.D.

and

Arlene M. Geist, M.A.

June, 1976

*Selections from the Seattle Community Accountability Program Crime Impact and 12-Month Recidivism Analysis have been excerpted for inclusion.

I N T R O D U C T O R Y C O M M E N T S O N :

"SEATTLE COMMUNITY ACCOUNTABILITY PROGRAM (FORMERLY YOUTH SERVICE BUREAU SYSTEM) CRIME IMPACT AND 12-MONTH RECIDIVISM ANALYSIS," June, 1976

One of the quasi-experimental approaches used in the CAP evaluation involves utilization of an actuarial table, based on analyses of 90,000 pre-project youths with police contacts, for predicting the expected recidivism rates of youths in the project. This approach is explained in the excerpt from the evaluation and some of the problems with it are identified. Although many sections of the original report are not included here, the reader should note the very careful explanation and justification given to each of the performance measures used in the original evaluation. The theory underlying the project approach for reducing juvenile delinquency is explained and is linked to each of the performance measures. In addition, the assumptions inherent in the use of these measures as indicators of project effectiveness are explained and apply to most evaluations of similar projects.

CRIME IMPACT

A. Evaluation Design

The goal of the Seattle Community Accountability Program (CAP), formerly the Seattle Youth Service Bureau (YSB) System, is to reduce juvenile crime in selected target areas of the City of Seattle. The implementation of the CAP's, in conjunction with Community Accountability Boards (CAB's), was designed to achieve this goal through both direct and indirect effects upon juvenile offenders. The direct, or primary, effect of preventing an offender from committing additional crimes was hypothesized to occur when individual youths were obliged to perform either monetary or community service restitution for their offenses. The indirect, or secondary, effect of preventing others from committing crimes was hypothesized to occur by locating accountability boards within CAP census tract areas; the accountability boards would deal with all (or as many as possible) of the juvenile offenders residing within those areas, regardless of where the actual offense may have occurred. It was assumed that the knowledge of such a program would become known to the youths in the CAP area and serve as a deterrent.

Since the program design involved the "treatment" of all juveniles residing in the bureau areas, the preferred evaluation design of assessing crime impact by randomly assigning youths to the accountability board process (experimental treatment) and the traditional criminal justice process (control treatment) was not possible. The evaluation design chosen consisted of a series of non-equivalent control group design comparisons (Campbell and Stanley, 1963), and comparisons of juvenile offenders' recidivism with actuarial predictions of recidivism (Youthful Offender Criminal History Survey Project, 1976).

To measure crime impact, three measures were chosen: individual youths' Seattle Police Department contact histories (a contact being equivalent to an adult primary, or major, charge); total number of juvenile contacts, by census tract of offenders' residence; and the reported occurrence of residential burglary, larceny and auto theft, by census tract, regardless of whether suspects may have been identified or arrested.

The reasons for choosing these particular measures are as follows: To assess the program's direct effect upon crime, the most logical measure is some index of treated youths' subsequent criminal behavior. However, the point at which this measure within the criminal justice system is made is a source of some controversy. Some suggest that to insure that those arrested are truly guilty, only those youths adjudicated guilty be counted. Others suggest self-

report is the only truly valid index. Data reported by Gold (1975) indicate that, based upon a self-report study, only 3 percent of juvenile crimes result in an actual arrest. However, other studies of self-report crime data raise serious questions regarding the accuracy and validity of such measures. Because of the cost factors and questionable reliability and validity of self-report measures, it was decided to deal with official criminal justice system data as a measure of recidivism. Keeping in mind Sellin's (1931) statement that "...the value of a crime rate for index purposes decreases as the distance from the crime itself, in terms of procedure, increases," police contact or charge data were chosen as the index of juvenile recidivism. This includes cases in which arrests initially were made and then investigated and released. In 1974, 9.2 percent of juvenile contacts within the City of Seattle were of this nature.

To assess the impact of the program's indirect effect, total juvenile contacts of youths residing in the CAP census tracts were chosen to be compared with the contact rate for non-CAP census tracts within Seattle. It was felt that, to the extent that the program had an effect of practical significance, it should be detectable on a census tract basis. The reason for choosing police contact data was the same as that given in the preceding paragraph.

The third measure, the reported number of residential burglary, larceny and auto thefts, was chosen to provide a relatively independent measure of crime, and to insure that the conclusions based upon police contact data were not misleading. Whereas arrest or contact data may represent as little as 3 percent of actual juvenile crime committed (Gold, 1975),crime victimization studies conducted both nationally and in Seattle (U.S. Department of Justice, 1975a, 1975b; Schram, 1973; Mathews, in preparation) indicate that residential burglary is reported in approximately 45-55 percent of victimizations; larceny, approximately 20-40 percent; and auto theft, approximately 70-90 percent of all victimizations. In addition, reports of crime occurrence are less susceptible to change due to changes in police procedure within the program area. That is, one might suggest that changes in police contacts within the CAP area may be caused by either decreased or increased activity in apprehending juveniles, rather than program effects. (There has been no known change in police manpower or activities in the project area that would substantiate such a suggestion.) However, it would be unlikely that the presence of the CAP accountability board system within various census tracts would be associated with a change in the reporting of crimes occurring within those areas.

The adequacy of the choice of this last measure requires two assumptions: first, that juveniles be involved in the commission of these crimes; and second, that these crimes be committed by local residents. If these assumptions can be met, then the three selected crimes should provide an independent measure to assess the combined general and specific deterrent value of the program.

An analysis of the age of individuals arrested and charged in 1974 for burglary, larceny and auto theft by the Seattle Police Department indicates that 75 percent of burglary, 69 percent of larceny and 78 percent of auto theft charges involve juveniles. Although these figures do not necessarily mean that a corresponding percentage of all such crimes are committed by juveniles, it none the less does reflect crimes which have a high degree of juvenile involvement (4,301 separate juvenile charges of the total 10,410 juvenile contacts in Seattle in 1974).

The second assumption, that crimes are committed by local residents, is substantiated by Turner (1969) and Mathews and Mobley (in preparation). Both of these studies measured the distance between juvenile offenders' places of residence and the location of the commission of an offender's crime. Both Turner (using 502 cases) and Mathews and Mobley (using 8,990 cases) found that over 50 percent of all juvenile crimes occurred within less than half a mile from juvenile residences.

The present crime impact report presents updated data on the Mt., Baker, Ballard-Fremont, and Southeast CAP's from September 1, 1974 through February, 1976 which was not included in the prior evaluation (April, 1976). This excludes those clients who participated only in the first year of operation of the Mt. Baker CAP. Serious time constraints precluded the inclusion of the updated recidivism data for these youths. In addition, only 11 of the clients available for follow-up (N=94) were heard by the Mt. Baker Accountability Board (CAB) during the first year. Since the major objective being tested with this program is the effect of the CAB process on recidivism, it was believed that the exclusion of only 11 additional cases would not substantially affect the recidivism analyses results.

3. Objective Three, Data Analysis

Objective three specified in the evaluation design was the following:

> Given a juvenile offender's participation in a CAP, significantly fewer numbers of youths will be shown to recidivate as compared with the predicted probability of recidivism.

Because the manner of program implementation did not allow random assignment of youths to experimental and control groups, which would be the preferred evaluation design, actuarial predictions of recidivism were used to create a "statistical" control group for comparison purposes.

As a measure of the extent to which the CAP reduces individual client recidivism, probability tables developed through the Seattle Law and Justice Planning Office Youthful Offender Criminal History Survey Project (1976) were employed. These tables, based on approximately 90,000 juvenile police contacts occurring in Seattle over a 20-year period, provide the probability of a given youth committing a subsequent offense, based on the age, race, sex, offense and number of prior offenses. For example, the probability of a black male, age 17, who has been contacted for a burglary which is his second police contact, being contacted for a third offense of any type within six months is .414. Predictions were made for 6-, 12- and 18-month followups. However, for the present CAP analysis, only the 6- and 12-month predictions were used, due to the small number of youths for whom followups of 18 months were possible. In addition to the probability of committing an offense, the tables include the average number of offenses committed by those youths who did recidivate within the 6-, 12- and 18-month followup periods.

Preliminary results on the accuracy of such actuarial predictions of recidivism for randomly selected, non-treated juvenile offenders living outside the CAP areas indicate that the predictions are not significantly different from actual recidivism for a 6-month followup. That is, they neither over- nor under-predict recidivism to any appreciable extent for a sample (n = 45) of the general population of juvenile offenders. More extensive validation efforts of the actuarial tables are currently in progress and will be reported in the next evaluation.

One possible concern regarding the use of this sort of analysis, given that CAP clients are not randomly entered into this program, is the possibility of a selection bias. This may have occurred at either one of two points in the referral process: (1) screening on the part of the King County Juvenile Court workers, resulting in only the least-likely-to-recidivate youths being referred to the CAP; and (2) screening on the part of the CAP staff, resulting in acceptance of only those youths least likely to recidivate.

To determine if some selective screening was occurring, a random sample of 44 youths residing in the CAP areas who were not referred to or accepted by the bureaus was selected. A comparison of 6-month actual and predicted recidivism for this group resulted in a non-significantly lower ($x^2 = 3.18$, $df = 1$, $p < .10$) actual recidivism than predicted. At this time, the analysis indicates that a selection bias is probably not occurring. However, given the relatively small number of youths in the sample, a more complete analysis involving more non-CAP youths within the CAP areas would be appropriate. Such an analysis is currently underway and will be reported in the next evaluation.

Objective Three was evaluated for the total program and by individual bureaus since sufficient numbers of juvenile recidivism records were available. While the three CAP bureaus saw 617 youths between September 1, 1974 and February 29, 1976, only 205 youths were suitable for the individual recidivism analyses. Only those youths who met the following criteria were included in the recidivism analyses:

1. Youths with prior offenses for whom no prediction was available in the Probability Tables were excluded from the 6-month individual recidivism analysis. This situation arose in about 24 percent

of the cases. Obviously, since the Probability Tables were based on actual contact records, all possible combinations of age, race, sex, offense and number of offenses could not be met.

2. Only those youths who had committed no prior offenses or had committed an offense within three months prior to CAP entry were included in the analysis of predicted vs. actual recidivism (after entry). Those youths who had committed no prior offenses had a 0 prediction of subsequent contact.

3. Only those youths who had at least six months followup from date of last contact (or from entry date in the case of no prior contacts) were included in the analyses.

a) Combined Bureau Recidivism Analysis

Using actual vs. predicted 6- and 12-month recidivism data, comparisons were performed first for the total CAP population available for follow-up. These data are presented in Table 3 for all CAP youths, all CAP youths who were heard by an Accountability Board (CAB) and all CAP youths who received services only (non-CAB youths).

Chi-square tests performed on the data in Table 3 indicate that significantly fewer ($p < .001$) youths recidivated within six months than would be predicted to recidivate. For the 12-month analysis, the results were marginally significant ($p = .06$). In addition, taking the youths who appeared before a CAB as a separate group for all three bureaus combined (see Table 4), a statistically significant difference ($p = .001$) was demonstrated between actual and predicted recidivism within 6 months. In this case, only 17 out of 150 CAB youths committed subsequent offenses within six months. The 12-month analysis was, again, marginally significant ($p = .08$) where 20 out of 82 CAB youths recidivated. When those youths having no CAB participation (i.e., Service Only youths) were considered as a separate group, no difference was found between actual and predicted recidivism. It appears, therefore, that the CAB experience (but not participation in CAP services only) is significantly related to reduced recidivism, at least during a 6- and 12-month followup period.

TABLE 3. CAP CLIENT RECIDIVISM FOR A SIX- AND TWELVE-MONTH FOLLOWUP

	Youths	# Youths Followed Up	# Actual Recidivators	# Predicted to Recidivate
6-MONTHS	All Youths	250	27	54.9
	CAB Youths	150	17	39.87
	Non-CAB Youths	100	10	15.07
12-MONTHS	All Youths	146	30	43.8
	CAB Youths	82	20	30.2
	Non-CAB Youths	64	10	13.57

TABLE 4. CAB CLIENT RECIDIVISM FOR A SIX- AND TWELVE MONTH FOLLOWUP

	CAB	# Youths Followed Up	# Actual Recidivators	# Predicted to Recidivate
6-MONTHS	Mt. Baker	38	4	9.6
	Ballard-Fremont	61	9	16.2
	Southeast	51	4	14.0
	Total	150	17	39.9
12-MONTHS	Mt. Baker	24	6	8.4
	Ballard-Fremont	33	9	12.4
	Southeast	25	5	9.3
	Total	82	20	30.2

An additional means to evaluate a reduction in recidivism is to determine if those youths who do recidivate commit significantly fewer offenses than would be predicted.

Table 5 presents data on the number of offenses per recidivator for all CAP clients who met the same criteria as above, with the additional requirement that they did, in fact, commit at least one offense during the follow-up period. This additional requirement insures that the number-of-offenses-per-offender analysis is statistically independent from the results of the number-of-recidivators analysis. Table 5 shows there was a significant difference ($p = .01$ and $p = .007$) between actual and predicted number of offenses when all youths who recidivated within 6 and 12 months, respectively, are considered. Analysis of individual bureau data was not performed because of the small number of youths who recidivated (See Table 4).

TABLE 5. NUMBER OF OFFENSES PER OFFENDER FOR CAP CLIENTS WHO RECIDIVATED WITHIN SIX AND TWELVE MONTHS

Youths	# Who Recidivated	Number of Offenses per Recidivator		One-tailed T-Test Probabilities
		Actual	Predicted	
(6 months) All Youths	27	1.93	3.15	$p = .01$
(12 months) All Youths	30	2.07	2.87	$p = .007$

b) Individual Bureau Recidivism Analysis

The analyses performed so far in relation to Objective Three (reduction of individual recidivism) have been relatively global; that is, all bureaus and all services combined. However, for formative evaluation or program modification purposes, it is desirable that relationships between the various service components or CAB appearances and reduced recidivism be demonstrated. In other words, it is necessary to determine the relationship between a dependent or criterion variable (in this case,

recidivism) and a set of independent variables (CAB and/or services received), if policy decisions regarding the exclusion of any service(s) from the CAP are to be made. Included in this analysis were only those youths who had either committed no pre-entry offense or had committed an offense no longer than three months prior to CAP entry. Excluded were youths who could not be matched in the probability tables. In addition, only those youths who had a follow-up of at least six months from date of entry or date of pre-entry offense were included in the analysis.

SECTION 4.F.

BURGLARY TASK FORCE EVALUATION*

By

Pierce County Law and Justice Planning Office

Steven R. Barlow

and

Elaine Kaufman

March, 1977

*Selections from the Burglary Task Force Evaluation have been excerpted for inclusion.

I N T R O D U C T O R Y C O M M E N T S O N :

"BURGLARY TASK FORCE EVALUATION"

The Burglary Task Force Evaluation Report is included here for several reasons. First, it demonstrates the value of a "project component" type of evaluation in which two strategies (fencing investigations vs. burglary investigations) of the same project are compared in terms of several performance measures. Second, the style of presentation is particularly good: The first page is a short summary (for busy decision makers). This is followed by a careful elaboration of the logic and rationale of the project strategies. The purpose and methods of the evaluation also were made clear. And, after a presentation of the results, the authors included an interesting discussion of the meaning, relevance, and implications of the findings for project operation.

The evaluation procedures also illustrate several important principles. Multiple performance indicators were used, including several productivity measures. The evaluators compared time-series data from the city with two other areas in order to examine the consistency in findings across the comparison. And, when the conclusions indicate little, if any, overall effectiveness of the project (incorporating both the fencing and burglary investigations) the authors attempt to determine why the project did not work.

Readers should be aware that many evaluation reports use a rather imprecise definition of "efficiency"--including this one. Efficiency should be measured as the ratio of outputs to inputs. The level of activities or the "outputs" per se should not be considered indicators of efficiency.

For example, the number of arrests is an indication of "activity level," but it is not an indication of efficiency or comparative efficiency. The number of man-hours required to produce an arrest is a measure of efficiency. The number of man-hours required to prosecute each $100 worth of value in a crime is an indication of efficiency.

In the BTF evaluation, it appears as if the number of man-hours was more or less equivalent across the burglary and fencing investigations, and more or less equivalent during the post-program year and the pre-program year. Thus, the use of activity levels and direct comparisons is not particularly biased, but the data would be better if more precise efficiency ratios were established, especially between the burglary versus fencing investigations.

BURGLARY TASK FORCE EVALUATION

The Burglary Task Force (BTF) is a specialized investigation unit whose primary focus is on stolen property trafficking (fencing). The emphasis placed on the investigation of fences has proved to be equally effective, but no more effective, than when investigations are restricted to burglars. Nor has the orientation to fences significantly changed the case profile or productivity of BTF detectives over what they accomplished during an earlier but equivalent time period. BTF has not served as a significant deterrent force to burglary in Pierce County. The limited success of the unit is partially attributed to a lack of resources. No provision was made for "buy money" in the BTF grant and consequently BTF investigators were forced to perform their job without one of the primary tools of the fence detective. Finally, BTF detectives emphasized investigation activities and did not meet objectives related to the development and dissemination of information describing fence operations to other law enforcement agencies.

The following is a report of the evaluation of a specialized investigation unit of the Pierce County Sheriff's Department; the Burglary Task Force (BTF). The report is divided into four sections: The introduction was drawn from the BTF grant to convey the original program concept as accurately as possible. The method section describes the evaluation design, while the results and discussion sections present the data analysis and interpretation of results, respectively.

INTRODUCTION

Problem

Law enforcement investigations in Pierce County note that there is no intrinsic profit in stealing property, and that the basic motivation for stealing is conversion of property to cash. While some stolen goods are easily sold to citizens careful not to ask questions about bargain prices, most are sold to fences who purchase and dispose of goods through legitimate business channels or outlets.

Intelligence sources indicate that more than thirty fences operate in Pierce County; that they handle goods stolen not only locally, but also from adjoining counties and states. These same sources report that goods stolen from within the area are disposed of by burglars from thirty minutes to two hours from the time of the theft.

During recent years, traditional methods of investigation by detectives have not been sufficient to long deter professional fences moving stolen goods. Better identification of these goods is one strategy for which positive results are anticipated. A second and equally important strategy has been identified through the Portland, Oregon LEAA High Impact Program. For a limited period in 1974, a special detail was established by the Police Bureau as a mission within the Impact Strike Program. The purpose of the detail was to focus exclusively on fences through use of enforcement; sophisticated, legal surveillance techniques; and accurate preparation and service of search warrants. The small, special detail arrested 102 persons during a six-month period and cleared 175 cases, recovering $93,000 in stolen property. Of the recovered property, the detail returned $87,000 worth to the victim for a phenominal return of 93%. Based on the initial results of this approach, the City of Portland and the Bureau of Police formed a full time special detail to work fencing operations in the area. Since the implementation of the unit funded through LEAA, extraordinary results have been obtained.

The BTF project addresses the following:

Nighttime, residential burglaries are easy to commit.

Disposal of property stolen in nighttime, residential burglaries is easy.

This project addresses both of the previously stated problems. Statistical analysis has provided the basis by which we can state that of residential burglaries, fewer than one out of five of the known burglaries are reported to the police, which likely means that fewer than one of ten burglaries are ever solved. In support of the second problem, we found that less than 10% of the property stolen in nighttime residential burglaries, as well as burglaries generally, is recovered.

This project, by using intelligence information gathered by the task force, should increase the arrest rate 10% above the 1974 burglary arrest rate. More specifically, the unit, by increasing the surveillance of known and suspected fences, will reduce the disposal of stolen property.

Project Intent

A. Goal

The goal of this project is to reduce the disposal of property stolen in nighttime, residential burglaries.

B. Objectives

1. Through exchange of ideas and techniques, develop new and more efficient approaches to burglary investigation.

2. Identify, investigate, and apprehend fences and persons who knowingly purchase or dispose of stolen property.

3. Determine if intrinsic interdiction of Pierce County fence activities will affect the incidence of commercial and residential burglaries.

4. Develop information regarding fence activities for dissemination to appropriate law enforcement agencies in the Pierce County area.

5. Recover, identify, and return stolen goods to the owner-victim.

6. Reduce burglary rates in the first year by 10%.

Project Implementation

1. Arrest three persons per month who buy, sell, receive, or possess stolen property.

2. Arrest or referral of three persons per month for burglary/theft offenses.

3. Investigation of five persons or locations each month where there is suspected buying, selling, receiving or possession of stolen property.

4. Preparation and service of three project related search warrants per month.

5. Increase the amount of stolen goods recovered by the Sheriff's Office by 5%.

6. Identify and return to victim 80% of stolen goods recovered.

7. Conduct four public education programs per year regarding property crimes and fencing activities in the Pierce County area.

8. Conduct one information seminar per month for appropriate law enforcement agencies in the Pierce County area regarding fencing investigations, search warrant preparation, identification and return of stolen goods.

9. Achieve a working knowledge of how each agency handles burglaries.

10. Standardize burglary reporting system for all appropriate law enforcement agencies in Pierce County.

11. Design and maintain a comprehensive and current record system composed of a burglary intelligence hot sheet, interagency burglary intelligence form and information currently available from Pierce County law enforcement agencies as well as state and federal agency methods.

12. Utilization of reliable informants and undercover officers to carry out covert investigation.

13. Maintain close working relationships and exchange information with department's intelligence unit.

Staffing will consist of two experienced burglary detectives assigned full time to the unit. One intelligence officer will be assigned to collect and supply all raw information pertaining to burglary and fencing operations obtained from street information. Funding for this officer is provided by departmental budget. The unit will report directly to the Chief Criminal Deputy who will assist in general strategy development and provide necessary command level liaison with other department units.

In addition, one Deputy Prosecuting Attorney experienced in the prosecution of burglary cases will be available by the Prosecuting Attorney's Office to assist the unit. This will involve the prosecutor and detective assigned to the unit working together during the evidence gathering stage and in preparing the cases for trial.

The key to solving crimes is investigation and the key to convicting arrested individuals is enhancing evidence. Due to the lack of resources and lack of consultation between detectives and deputy prosecutors, many crimes remained

unsolved. It is believed that many crimes are committed by a few individuals and a thorough job of investigation and continuous follow-up consultation between detectives and prosecutors should generate better cases against target offenders. The deputy prosecutor will:

1. Assist in answering legal questions during investigation.
2. Advise and participate in the securing and service of search warrants.
3. Handle all phases of cases in court during prosecution.
4. Advise the investigator as to building a prosecutable case.
   a. Avoid legal technical defense.
   b. Save investigative time by focusing the investigation towards effective prosecution.
5. Be available for consultation on any of the above problems on a twenty-four hour basis, if necessary.

Evaluation Purpose

The evaluation was intended to investigate the effectiveness of the program concept and its impact on burglary perpetrated in Pierce County. It is an interim report of BTF activity during the period of 3-1-76/12-31-76. The purpose of the evaluation was to provide Law and Justice Committee members with objective information with which to determine future funding of BTF. In addition to supporting funding decisions, the evaluation was conducted to provide management feedback to the project to facilitate program planning for future BTF operations.

METHOD

A. Criteria

1. Measures of effectiveness

   a. Technique effectiveness

      BTF detectives use two different techniques of investigation; one which is directed at the burglar and the other which focuses on fencing operations. The following indices of technique effectiveness will be employed:

      1. Economic loss
      2. Number of arrests

   b. Further analysis of the effectiveness of BTF was made by comparison of the investigation cases closed by the BTF detectives during the BTF project period to the burglary investigation cases closed by the same detectives during the year prior to the start of BTF. The

comparison was made on the basis of:

1. Economic loss
2. Number of arrests

2. Impact evaluation

a. Actual burglaries reported in Pierce County during the 12 months just prior to the BTF was compared with actual burglaries which were reported during the project period. Change in the incidence of burglary in Pierce County was compared to the change in burglary reported by a matched comparison group.

b. Arrests for burglary in Pierce County during the 12 month pre-project period were compared with arrests made during the project period.

3. Measures of efficiency

A process evaluation was conducted to monitor:

a. Time spent in all grant activities
b. The number of public education programs
c. The number of interagency information exchange sessions
d. The monetary value of recovered stolen property returned to victim
e. The number of search warrants prepared and served.

B. Data

1. Data collection

a. Method

Data regarding each investigation technique and grant activity was collected with a daily activity form which accounts for the detectives' time. All other information was collected from the detectives' logs, case records, or the Sheriff's Office management information system.

b. LJPO evaluation responsibilities

The program evaluator of the LJPO monitored implementation and data collected, and conducted the final analysis of the BTF evaluation.

RESULTS

Table 1 presents summary characteristics of investigations closed prior to 1-1-77. Four types of cases are described:

1) burglary investigations conducted under the guidelines of the BTF grant.

SUMMARY OF CLOSED[1] INVESTIGATIONS

Table 1

SUMMARY CHARACTERISTIC	N=28 BTF Burglary Investigations x	n	N=10 BTF Fence Investigations x	n	N=38[2] All BTF Closed Investigations x	n	N=36[3] Burglary Inv. Closed in 1975 x	n
Median number of days open	20	25	2	7	17	32	N/A	
Median manhours spent on case	16	12	22	8	19	20	N/A	
Total Est. Economic Loss	34,832	28	37,646	10	72,478	38	35,952	36
Median Value of Economic Loss	518	28	1081	10	682	38	260	36
Estimated Value of Recovered Property	19,506	23	34,946	6	54,452	29	6559	31
Median Value of Property Recovered	332	23	1975	6	650	29	100	31
Total Number of Arrests	38	28	14	10	52	38	53	36
Median Number of Arrests	1	28	1	10	1	38	1	36
Number of Adults Arrested	30	28	13	10	43	38	24	36
Number of Juveniles Arrested	8	28	1	10	9	38	29	36

1 = 56%
2 = 93%
3 = 75%
4 = 18%

1. Closed as of 12-31-76
2. Includes 12 commercial burglary investigations
3. Includes 14 commercial burglary investigations

2) fencing investigations conducted under the BTF grant.

3) all BTF cases, regardless of whether they were fence or burglary investigations.

4) burglary investigations conducted by BTF detectives during the year just prior to the start-up of BTF (4-1-75/12-31-75).

Some of the more distinguishing features of Table 1 are:

1) BTF fence investigations are open for a much shorter time than are BTF burglary cases.

2) BTF fencing cases require more manhours of investigation than do BTF burglary cases.

3) The median economic loss estimated by investigators was twice as large in BTF fencing cases as estimated for BTF burglary cases and just over four times as much as pre-BTF burglary cases.

4) The median value of property recovered as a result of fencing investigations was six times greater than in BTF burglary investigations and twenty times greater than in pre-BTF burglary cases.

5) The percentage of the suspected economic loss that was recovered was:

a)	BTF burglary cases	56%
b)	BTF fence cases	93%
c)	Pre-BTF cases	18%

6) There was no difference in the number of arrests made under any of the four categories described in Table 1. However, only 17% of BTF arrests were juveniles while 55% of pre-BTF arrests were juveniles.

Table 2 lists the characteristics of cases at the time they were filed in court. The same four categories are described in Table 2 as were presented in Table 1. The highlights of Table 2 are:

1) The median economic loss charged in BTF fencing cases was 2.35 times greater than BTF burglary cases and 3.43 times greater than pre-BTF burglary cases which were filed in court.

2) The median number of victims, counts, and defendants is equivalent among all categories of Table 2.

3) The conviction rate is very similar between BTF and pre-BTF cases.

4) The sentencing strategy is quite similar between BTF and pre-BTF cases. This observation is based on the percentage of cases where restitution was ordered or fines were imposed and the percentage of defendants sentenced to jail.

SUMMARY OF PROSECUTED CASES
Table 2

SUMMARY CHARACTERISTIC	N=26 BTF Burglary Case x	n	N=7 BTF Fence Case x	n	N=33 All BTF Closed Cases x	n	N=16[3] Burglary[4] Cases Closed in 1975 x	n
Median Economic Loss	1275	22	3000	7	1600	29	875	16
Total Economic Loss	56,000	22	41,850	7	97,902	29	39,654	16
Median Victims	1	25	1	7	1	32	1	16
Total Victims	27	25	8	7	35	32	26	16
Median Counts	1	25	1	7	1	32	1	10
Total Counts	35	25	12	7	47	32	21	16
Median Defendants	1	25	1	7	1	32	1	16
Total Defendants	39	25	10	7	49	32	23	16
Conviction Rate	88%	17	N/A[2]		90%	19	93%	15
% cases; Restitution Ordered	27%	15			25%	16	21%	14
Median Restitution Ordered	250	3			250	3	915	3
Total Restitution Ordered	1468	3			1468	3	3621	3
% Cases Fined	27%	15			31%	16	29%	14
Median Fine	116	4			75	5	225	4
Total Fines	546	4			621	5	757	4
% Defendants Sentenced	100%	27			96%	28	100%	18
Median Sentence	3 yrs	27			3 yrs	27	15 yrs	18
% Deferred, Suspended, Probation	67%	18			64%	18	44%	8
% Incarcerated	33%	9			32%	9	56%	10
Median Incarceration	3 yrs	9			3 yrs	9	3.5 yr	10

1. Cases prosecuted as of 1-15-77
2. The remaining characteristics are founded on cases which have been disposed in court. Only two fencing cases have reached this status, which is insufficient to provide summary characteristics.

3-4 (over)

3. Represents adult arrests only. Eight adults were not prosecuted because:

   4 had charges dropped against them
   1 could not be extradited
   1 was already jailed
   2 unknown

4. Represents burglary cases closed by BTF detectives during the year prior to the start of BTF.

CONTINUED

5 OF 7

The primary difference between the dispositions of BTF and pre-BTF cases was in the percentage of defendants who were actually incarcerated. Sixty-four percent of the BTF defendants received a deferred, suspended or probation sentence while only 44% of pre-BTF cases were so disposed. Consequently, only 32% of the persons arrested by BTF detectives who were subsequently convicted spent time in jail. Fifty-six percent of pre-BTF defendants were incarcerated.

A simple tabular comparison of case medians does not help decide whether or not a difference is anything more than a random, chance occurrence. Therefore, a statistical analysis of information derived from BTF and pre-BTF investigations and court cases was conducted. The first concern of this analysis was whether the economic loss suspected in BTF burglary cases was different from that of BTF fence cases. A t-test between the economic loss charged in BTF burglary cases filed in court and BTF fence cases filed in court resulted in $t = 1.55$, $p > .05$. No difference between BTF burglary and fence investigations can be inferred from this result. When the test is applied to the economic loss suspected during the investigation, the result is $t = 1.58$, $p > .05$. Again, no difference was found between BTF burglary and BTF fence investigations.

The number of arrests made as a result of an investigation is a second indicator of the significance of an investigation of an economic crime. Economic loss provides some estimate of the magnitude of the crime while the number of arrests made reflects the potential impact of the investigation as well as the size of the case. A t-test of the difference between the number of arrests made in BTF burglary cases and BTF fence cases showed no difference, $t = .45$, $p > .05$.

Another inquiry was directed at the difference between the investigations that resulted from the BTF program and those that resulted from the work of the same detectives during the year prior to BTF. A t-test of the difference between the economic loss estimated during the BTF investigation and the loss estimate for pre-BTF investigations demonstrated no difference, $t = .99$, $p > .05$. Similarly, no difference was found between the economic loss of BTF cases filed in court and pre-BTF cases filed in court, $t = .55$, $p > .05$. In regard to the number of arrests made, no difference was found between BTF and pre-BTF investigations, $t = .758$, $p > .05$.

Thus far, the analysis has been directed at the effectiveness of BTF services. But what of their impact on the target crime of burglary? To answer

this question, the change in the number of burglaries which occurred in Pierce County over the years of 1974, 1975 and 1976 was compared to the change in burglary experienced by the City of Tacoma during the same time period. If BTF has served as a deterrent force against burglary, it would be expected that the rate of increase in burglary established during 1974 and 1975 would be reduced significantly in 1976. Furthermore, this reduction should be unique to Pierce County when it is compared to a similar region which does not have a specialized fence investigation unit. If other areas are experiencing the same decline in burglary as Pierce County despite the fact that they don't have a BTF, the change in the burglary trend must be attributed to something other than BTF.

The City of Tacoma was selected for comparison to Pierce County because:

1) Tacoma has experienced a similar number of burglaries as has Pierce County.

2) The rate of increase in burglary is similar between Tacoma and Pierce County.

3) With the exception of BTF, Tacoma has all of the same burglary reduction programs as Pierce County.

4) Tacoma's proximity to Pierce County increases the likelihood that non-law enforcement variables which affect burglary are much the same for the two jurisdictions.

To conduct the impact analysis, the number of actual burglaries which occurred each month of 1974, 1975 and 1976 was determined for the City of Tacoma and for Pierce County. A linear trend analysis of 1974 and 1975 burglaries was then conducted. Given the past trend, the amount of burglary which was expected to occur each month of 1976 was predicted. The impact analysis consisted of a comparison of actual and predicted burglary between Tacoma and Pierce.County. The comparison was made using a zxz analysis of variance where predicted and actual burglaries represented the rows and Tacoma and Pierce County represented the columns. The above described method of impact analysis was employed to:

1) isolate BTF as an independent variable

2) consider the history of burglary in Pierce County.

The results demonstrated a significant row effect, $F = 54.4$, $p < .05$; but no column effect, $F = .007$, $p < .05$; and no interaction, $F = 1.88$, $p > .05$.

These results are interpreted to mean that there was a significant difference between the actual and predicted burglary for 1976. However, there was no difference between Tacoma and Pierce County and the magnitude and direction of the effect of the different jurisdictions must be assumed to be the same. The increasing trend in burglary was stopped, but it was stopped equally for both Tacoma and Pierce County. Therefore, the change in burglary demonstrated for 1976 cannot be inferred as the result of the impact of BTF on burglary in Pierce County.

A similar technique was used to determine whether the number of burglary arrests made by the Pierce County Sheriff's Office (PCSO) increased during 1976. Such an increase would be expected as a result of the additional resources provided PCSO by BTF. A linear trend analysis was computed for the actual arrests made by PCSO each month of 1974 and 1975. The trend analysis served as a basis for predicting the burglary arrests expected for each month of 1976. The actual number of arrests was subtracted from the predicted number of arrests and a matched pair t-test applied to the difference scores. The result was $t = -2.99$, $p < .05$.

DISCUSSION

The primary source of information pertaining to fence operations is the burglar or other such informant. For this reason, burglary investigations and fence investigations are very much interrelated. The BTF detective investigates and apprehends the burglar to obtain information which will lead him to a fence. Consequently, 64% of BTF closed cases were burglary investigations, while 23% were strictly fence investigations. For every fence investigation, 2.8 burglary investigations were conducted.

One of the more fundamental questions of this evaluation is whether the additional orientation to fencing is more beneficial than if BTF had only investigated burglaries. Statistically, there was no difference between BTF burglary and fence cases. Although the average economic loss suspected in fence cases exceeds that of burglary cases, the difference is not significant at the .05 level. It should be noted, however, that this difference did exceed the .1 level of significance. The mean economic loss estimated for fence cases tended to be larger than for burglary cases, but one cannot be highly confident that this difference is anything more than a chance occurrence.

Regardless of the mean differences between fence and burglar cases, the latter resulted in greater total productivity. Burglary investigations were found to exceed fence investigations in total economic losses charged in court, in the number of victims served, and in the total number of defendants convicted. From the standpoint of total productivity, it appears that many burglary investigations produce greater results than a few fence investigations. The ratio of burglary to fence cases is not expected to change, either. It is more difficult to obtain information about fencing operations and there are simply fewer fences.

Comparison of BTF burglary and fence cases leads to the conclusion that investigation of persons who traffic stolen property will not enhance either the magnitude of individual cases or total productivity.

There seems to be little advantage for BTF to investigate fences other than the ethical motivation to investigate all types of criminal acts. On the other hand, there doesn't appear to be any disadvantages, either. The short, intensive fence investigation consumes but a few more manhours than does the burglary investigation and lasts for only a couple of days. Therefore, investigative resources are not disproportionately consumed by fence investigations. Relative to the number of manhours spent, had the 10 fence investigations not been conducted, only 12.5 burglary investigations would have been made in their place. Consequently, investigation of strictly burglary cases would have been no more effective nor productive than had BTF operated as they did.

The effectiveness of BTF fence investigations was measured in terms of magnitude rather than impact. If, however, fence cases serve as a deterrent to burglary, then the size of the case is a secondary and surely less important consideration. By taking away the fence, the burglar may be thwarted in his/her attempt to turn stolen goods into cash. Elimination of the monetary reward for theft may reduce the motivation to burgle. As a result, the incidence of burglary in Pierce County may be reduced. This argument makes it clear that the primary benefit of arresting fences should be measured in terms of crime impact rather than the suspected amount of economic loss, number of arrests, etc. Actually, the former represents the main thrust of the BTF grant.

Burglary has increased every year since at least 1972. In 1976, that increasing trend was stopped. Was this a manifestation of the impact of BTF? Comparison of Pierce County to the City of Tacoma demonstrated an equivalent burglary decrease in Tacoma, yet TPD does not have a burglary task force. This result makes it difficult to believe that BTF had an impact on burglary committed in Pierce County. It can, however, be argued that the arrest of fences in Pierce County had a spillover effect on the City of Tacoma. Fences do not restrict their activities to political subdivisions. Therefore, by arresting a fence in Pierce County who had also operated in the City of Tacoma, BTF may have had an impact on burglary in Tacoma. To investigate the spillover hypothesis, a second comparison was made to Spokane County. Spokane, like Tacoma, has the same burglary reduction programs as Pierce County except for a specialized fence investigation unit. Burglary in Spokane County was:

1973	1974	1975	1976
4550	4957	5010	4826

The increasing trend of burglary was stopped in 1975 and the incidence of burglary in Spokane County actually declined in 1976. The comparison of burglary among Pierce and Spokane counties and the City of Tacoma leads to the conclusion that the change in burglary during 1976 that was found for Pierce County was not a result of BTF.

An additional perspective of BTF services was made relative to a burglary unit which did not investigate fencing operations. The work of BTF detectives during the year prior to the start-up of BTF was selected for this purpose. Comparison of all BTF cases with pre-BTF cases revealed few differences. The average economic loss for individual cases investigated and prosecuted during 1975 was found to be statistically equivalent to BTF burglary and fence cases. The average number of arrests made per case was also equivalent for the two periods. The cases did not differ either in the number of victims or the number of defendants prosecuted. Finally, the conviction rate achieved through the investigative efforts expended by the BTF and pre-BTF investigative units was found to be the same.

The major difference between BTF and pre-BTF cases is in the number of juveniles investigated and/or arrested. Over half of the pre-BTF arrests were juveniles, while only 17% of BTF arrests were of juveniles. This may account for the greater, more than double, total economic loss estimated for BTF investigations over what was suspected in pre-BTF investigations (see Table 1). When the same comparison[1] is made relative to the economic loss charged in court, which excludes juvenile cases, the difference is nearly eliminated. (See Table 2). This alludes to the greater magnitude of adult burglaries with respect to juvenile burglaries. A second difference between BTF and pre-BTF cases is in the percentage of property that was recovered. Seventy-five percent of the estimated economic loss of BTF cases was recovered while only 18% of the property reported stolen in pre-BTF cases was recovered. From the evidence presented in this report, it appears that BTF provides a special investigation unit which orientates toward adult burglars and recovers a high percentage of the suspected stolen property. Otherwise, the special emphasis on stolen property trafficking

[1]The comparison was made by doubling the economic loss of pre-BTF cases so as to approximate the number of BTF cases.

has not significantly changed either the productivity or the case profile relative to pre-BTF endeavors.

Why has the success of BTF been so limited? Discussion with the BTF detectives and management has provided some insight. The most effective technique of investigating fencing operations is for the detective to establish a criminal relationship with the fence. That is, the detective must convince the fence that he is either a burglar who wants to dispose of stolen goods or a buyer of stolen goods. The criminal interaction with the fence allows the detective to develop evidence against that person and to obtain information which may uncover other fencing operations.

To buy from a fence, the detective must have what is referred to as "buy money". Buy money is one of the primary tools of the fence investigator. Yet, the BTF grant did not make such a provision. BTF detectives have had to operate without buy money and, thus, have been severely hampered in their endeavors to investigate fences. Most simply, BTF detectives have been asked to do a job without the appropriate tools. It is the recommendation of this evaluation that some provision for buy money be made before continued BTF funding can be considered.

The objectives of BTF were twofold. The first type of objective related to the deterrence of burglary through the apprehension of fences and to the recovery of stolen property. These have been discussed. The second type of objective addressed by BTF conveyed their desire to develop information regarding fence activities for dissemination to appropriate law enforcement agencies in Pierce County. The activities which were proposed to meet this objective were:

1) Fencing information exchange sessions with other agencies. Reference to Table 3 shows that 2% of BTF time was spent in eight information exchange sessions with valley peace officers.

2) Public education. One public education session was held for used car dealers to consume less than 1% of BTF time.

3) Standardization of burglary reports. No time was given to this activity.

4) Development of interagency hotsheets. No time was given to this activity.

It is quite clear that BTF did not emphasize the information development and dissemination objective. This finding may be a function of resources. To expect two detectives to have an impact on burglary through the apprehension of fences and to also develop a standardized burglary information system for all law enforcement agencies in Pierce County may have been unrealistic. Given the resources of two detectives, one or the other type of objective may be addressed, but not both. It is the recommendation of this evaluation that the objectives stated in the BTF grant be revised to reflect the actual activities of the BTF detectives.

A final comment on the effect of BTF is made with regard to the increase in arrests made by PCSO. Arrests for burglary increased significantly during 1976 over what was expected from the arrest trend established by PCSO in 1974 and 1975. This is inferred to be a result of BTF. Quite simply, the BTF grant provided additional resources to PCSO. Therefore, the number of arrests made during 1976 increased as a function of the increased resources given PCSO for the investigation and apprehension of burglars and fences.

In conclusion of this report, BTF has been found to provide additional resources to PCSO for burglary investigations. However, the unit has not had a significant impact on burglary in Pierce County. Investigations which have been focused on stolen property trafficking have proved to be no more effective than burglary investigations. This may be a function of the unit's lack of "buy money". It is recommended that the provision of an additional resource for buy money be a primary consideration in the decision to continue BTF funding. A further recommendation was made to redefine BTF objectives to reflect the actual activities of BTF detectives. Currently, they are not working toward information development and dissemination objectives.

SECTION 4.G.

BURGLARY PREVENTION TEAM EVALUATION*

By

City of Spokane
Law and Justice Planning Committee

JoAnn Ray, M.S.W.

April, 1978

*Selections from the Burglary Prevention Team Evaluation have been excerpted for inclusion.

INTRODUCTORY COMMENTS ON:

"BURGLARY PREVENTION TEAM EVALUATION"

This report contains an excellent description of the problems faced by an evaluator when the project was not operated in such a way that it could be evaluated. These problems include:

(1) The project activities were so diversified across types of burglaries, prevention strategy, areas of the city, and start-up time, that the independent variable (e.g., the "project") could not be clearly identified and the dependent variable (e.g., its target) could not be isolated enough to prevent any but the most rudimentary types of analysis.

(2) A large number of other new strategies were implemented within the same areas at about the same time. Even though these had somewhat different targets, all of the projects' targets tended to overlap in relation to type of burglary, prevention strategy, area of the city, and start-up time.

With these constraints, the only way to assess project effectiveness in crime reduction was to examine the impact on burglaries for the entire city.

The interrupted time series analysis demonstrates the difference in conclusions that would be reached if a pre/post design had been used instead.

EXCERPTED AND EDITED

BURGLARY PREVENTION TEAM EVALUATION

SUMMARY AND DISCUSSION

Grant Description

Burglary was defined by the Law and Justice Planning Committee in 1975 as Spokane's Number One Crime. This grant proposed to reduce burglaries in Spokane through prevention. Non-residential burglaries were targetted. Methods specified in the grant include: security check visitations of commercial establishments; promoting worthwhile crime prevention programs; explanation of anti-intrusion devices; public education through presentations and media; study of burglary trends; and study of legislation.

The evaluation criteria selected to measure the project's effectiveness include: change in the rate of non-residential burglary, change in the rate of no-force non-residential burglary, change in the monetary loss due to non-residential burglary, and the percentages of businesses who implement the officer's suggestions.

Constraints

No standards to determine project usccess were established prior to project implementation. It is, therefore, impossible to state whether the project reached the anticipated reduction in burglary, number of completed site visits, etc.

Because of the widely diversified activities of the Burglary Prevention Team, it is difficult, if not impossible, to isolate the impact of the program. The Burglary Prevention Team's primary responsibility has been non-residential burglary prevention; however, some of the activities,

especially through the media, have been related to residential burglary and other crimes also.

The large number of strategies utilized both locally and statewide increase the difficulty in assessing the effectiveness of this project. The local programs, Neighborhood Watch and Juvenile Court Burglary Reduction, while not specifically targetting non-residential burglary, may have an indirect effect upon non-residential burglary. The prevention information disseminated by the Washington Crime Watch Program from the Washington Attorney General's Office has been widespread.

Because of the diversified activities of the Project Staff, the possible impact of other projects and the lack of objective standards, it is impossible to draw any valid conclusions as to the effectiveness of the program. Information is presented as an indicator of the project's progress.

Findings

Both non-residential and residential burglary experienced an 18.2% decrease from 1976 to 1977. Although this decrease appears to be substantial, interrupted time series analysis indicates that there is no statistically significant difference in the short or long term change of the burglary rates during this time period.

An analysis of the trends of burglary from 1971 through 1977 indicates that non-residential and the total city burglary rates have shown a decline since 1975.

The percentage of no-force, non-residential burglaries estimated from a sample, changed two percent--from 20% in 1976 to 22% in 1977.

Although the overall loss from non-residential burglaries has decreased from $336,449 in 1976 to $317,573 in 1977, the average loss per burglary has increased from $337 to $389.

There were two-thirds of the business representatives interviewed who indicated that they had implemented the officer's suggestions for the improvement of business security. An additional one-quarter of the respondents indicated they intended to do so. Most often mentioned implementations included: improving locks, installing alarms, adding lighting, improving outside security and securing windows.

Records kept by the project staff indicate that 470 security checks were completed in 1977. There were 41 speaking engagements, 22 television presentations, 14 radio talk shows, 47 pre-recorded radio spots, six newspaper articles, and two magazine articles, presented to the Spokane residents on crime and burglary prevention during the year as a result of the project. The Burglary Prevention officers were involved in drafting a false alarm ordinance in cooperation with personnel from the alarm companies. The team has maintained a pin map and analyzed trends in the city's burglary problem.

The sample of business representatives who were interviewed expressed positive reactions to the officers' visit. Almost 100% of the respondents were favorably impressed with the officers' visit and thought the program would improve police/business relationships. Two-thirds of the business representatives had discussed the officers' visit with business associates.

GRANT REVIEW

Situation/Problem

Burglary was defined by the Law and Justice Planning Committee in 1975 as Spokane's Number One Crime. Reported burglary rates in the City of Spokane had increased 43% from 1970 to 1975 and clearance rates for burglary varied from 11% to 13%.

As a strategy to reduce burglary by increasing the arrest and conviction rate, the Burglary Reduction Team (BURT Team) was implemented by the Police Department in 1976. The grant provided funding for two officers who were assigned on a selective basis to immediately collect evidence, provide surveillance, and gather related information from victims and witnesses.

A monitoring report of the BURT team was completed by the Evaluator in the Fall of 1976. Concern over the lack of project success was expressed and a meeting was held with the project director and Law and Justice Office planner. At the meeting it was thought that a team or two officers would be more successful in impacting burglary through prevention rather than apprehension. The grant proposal for the second year was substantially rewritten.

During the second year the project's intent was to reduce burglaries in Spokane through prevention. Non-residential burglaries were targetted. This project attempted to impact the problem of burglary in a pro-active way, by working with the business community in an effort to better protect their facilities from the burglar.

GOALS:

1. To achieve a notable decrease in the level of non-residential burglaries in the City of Spokane.
2. To realize a reduction in the dollar amount of loss suffered by burglary victims, even above that amount in proportion to the reduced number of burglaries.

OBJECTIVES:

1. Given the intensive crime prevention efforts of the Burglary Prevention Team, there will be a decrease in the number of non-residential burglaries reported in 1977 over 1976.
2. Given the intensive crime prevention effort of the Burglary Prevention Team, there will be a decrease in the average monetary loss per case per non-residential burglary in 1977 over 1976.

IMPLEMENTATION

1. Conduct security check visitations to commercial establishments in the City of Spokane, based on requests and as shown by need.
2. Promote worthwhile crime prevention programs, e.g., Operation Identification.
3. Thoroughly explain anti-intrusion devices to business managers, e.g., alarm systems, locks, seals.
4. Appear before groups with displays, films, and make speeches and answer questions.
5. Perform burglary analysis tasks and provide this service to line personnel.
6. Study appropriate crime prevention legislation.

7. Respond to selected reported burglaries to determine problems in the community.

BUDGET

LEAA	$35,320.00
STATE	1,962.00
LOCAL	1,962.00
TOTAL	$39,244.00

EVALUATION DESIGN

I. The effectiveness measurements of the evaluation design consist of a pre/post comparison of the following:

1. The rate of non-residential burglary in the City of Spokane.
2. The percentage of no-force, non-residential burglaries in the City of Spokane.
3. The average monetary loss per non-residential burglary in the City of Spokane.
4. The percentage of business persons who have implemented the officer's suggestions for improvement of business security.

II. The efficiency measures of the project include a monitoring of tasks:

1. Number of speaking engagements, people reached, and organizations contacted.
2. The number of security checks completed.
3. The Burglary Prevention Team's activities in studying the need for legislations.
4. The Burglary Prevention Team's activities in burglary analysis.
5. Other activities of the Burglary Prevention Team.

[illegible] No standards to determine project success were established prior to project implementation. An analysis of the change in the rate of burglary has been completed using interrupted time series analysis. Other analysis is completed in percentages and efficiency measures are presented in raw numbers.

FINDINGS

I. EFFECTIVENESS MEASURES

A. A comparison of the number of non-residential burglaries for the years 1975 and 1976 to 1977.

Non-residential burglaries have decreased 10% from 1975 to 1976 and 18% from 1976 to 1977, or a total of 27% from 1975 to 1977.

1975	1113	92.8 per month
1976	999	88.3 per month
1977	817	68.1 per month

A comparison of the rates of residential burglaries for the years of 1975, 1976, and 1977 indicate an 18% decrease in 1977 over 1975 and 1976.

1975	2324	193.7 per month
1976	2335	194.6 per month
1977	1911	159.2 per month

It can be noted that there was an identical decrease of 18.2% in the rates of residential and non-residential burglaries from 1976 to 1977.

B. In order to determine whether there was a significant short or long change effect on the rate of burglary, an interrupted time series analysis was also completed. Interrupted time series analysis projects (forecasts) from the values of the pre-data the expected value of the post period. The actual observed data is compared with these theoretical expected values. If the intervention has made a significant impact in either a short term or a long term change, this comparison will be statistically significant.

Interrupted time series analysis was completed for the following: non-residential burglary, residential burglary and total city burglary.

The intervention point used in this analysis was January 1, 1977, the beginning of the Burglary Prevention Team's activities. Monthly UCR burglary rates for 1976 were used as pre-data and monthly UCR burglary rates for 1977 were used as the post data.

The analysis indicates that there was no significant short or long term change in the non-residential, residential, or combined burglary rates in the 1977 as compared to 1976.

As the effects of this projects are cummulative, it is possible that the change in the burglary rates would not be apparent as soon as January 1, 1978. Additional interrupted time series analysis using February, March, April, etc. as the intervention point will be completed at a later date.

...

SPOKANE BURGLARY

	1971	1972	1973	1974	1975	1976	1977
Total	2645	2686	3288	3485	3437	3333	2728
Residential	1801	1595	2161	2241	2324	2335	1911
Non-Residential	849	1091	1127	1244	1113	999	817

2. A comparison of the percentage of no-force, non-residential burglaries for 1975 and 1976 to 1977.

A sample of February, May, August and November non-residential burglaries was used. The percentage of no-force non-residential burglary has changed only 2% from 1976 to 1977.

	1975	1976	1977
Force	74%	76%	71%
No Force	21%	20%	22%
Unknown	5%	5%	7%

3. A comparison of the monetary loss per non-residential burglary for 1975 and 1976 to 1977.

Although the overall loss from commercial burglary has decreased the past two years, the average loss per burglary has increased.

	1975	1976	1977
Total Loss from Non-residential Burglary	$361,852	$336,449	$317,573
Average Loss per Burglary	352.11	336.79	388.71

4. The percentage of business persons who have implemented the officer's suggestions for improvement of business security.

A sample of 68 business persons who have had a security check completed, indicates that 67% of those who were given suggestions implemented them. An additional 24% of the respondents indicated that they intended to do so in the future. (For more detail, see next page.)

VERIFICATION CALLS

A sample of businesses in which a security survey was completed the previous month was called to determine their attitude toward the officer's visit and whether they implemented suggestions made by the officers. A total of 68 calls were completed during the year.

Almost 100% of those surveyed had a positive reaction to the officer's visit. Comments included: "informative", "excellent", "helpful", "thorough". Only one respondent was negative--because the officers spoke to an employee rather than the manager.

The business representatives stated they initiated the officer's visit in 75% of the cases and the police officer in 25% of the cases.

Almost 90% of the persons surveyed indicated that the officer had made suggestions for improvement and 9% of the businesses were already adequately secure.

Of those given suggestions by the officers, two-thirds stated that they had implemented the officer's suggestions and another one-fourth stated that they intended to do so.

Most frequent responses to the question of what measures had they taken to improve their business security were: improved locks, installed alarms, added or changed lighting, improvements to the outside security and securing windows.

Almost 100% of those business persons surveyed indicated they thought that the program would improve police/business relationships.

Just over two-thirds of the business representatives sampled indicated that they had discussed the officer's security survey with business associates.

BURGLARY PREVENTION TEAM
VERIFICATION CALLS

1. What was your impression of the officer's visit?

Positive	67	99%
Negative	1	1%

2. Who initiated the visit?

Police	17	25%
Business	51	75%

3. Did the officer make any suggestions of ways in which you might improve your business security?

Yes	60	88%
No	2	3%
No (already adequate)	6	9%

4. Have you made use of these suggestions?

		Percent of Total	Percent of Those Given Suggestions
Yes	39	57%	67%
No	19	28%	33%
Not Applicable	8	12%	
No Information	2	3%	

5. If yes, what improvements? (multiple answers possible)

		Percent of Total	Percent of Those Given Suggestions
Improved Locks	19	28%	33%
Installed Alarms	9	13%	16%
Added or Changed Lighting	7	10%	12%
Improvement to Outside (gates, fences)	6	9%	10%
Secured Windows	5	7%	9%
Changed Method of Handling Money	3	4%	5%
Others	6	9%	10%

6. If no, why not?

		Percent of Total	Percent of Those Given Suggestions
Plan to do so	14	21%	24%
Other	3		
Question Not Asked	2		

7. Do you think this program will improve police/business relations?

Yes	66	97%
No	2	3%

8. Have you discussed the officer's visit with any of your business associates?

Yes	46	68%
No	22	32%

Percentages may not equal 100% due to rounding of numbers.

SECTION 4.H.

DRIVING WHILE INTOXICATED (DWI) IMPACT GRANT EVALUATION*

By

Northwest Regional Council
Washington State Traffic Safety Commission

Stuart Readio

November 10, 1977

*Selections from the DRiving While Intoxicated (DWI) Impact Grant Evaluation have been excerpted for inclusion.

I N T R O D U C T O R Y C O M M E N T S O N :

"DRIVING WHILE INTOXICATED (DWI) IMPACT GRANT"

The driving while intoxicated prevention (or "countermeasures") project presented the evaluator with several problems, many of them similar to those encountered in all field evaluations where there is no randomly assigned control group, no comparison group or area that can be used, and only monthly (aggregated) data are available (no case-by-case data). Much of the analysis relied on interrupted time series but the evaluator incorporated several interesting analysis techniques and strategies into this design.

One of these (pp.27-28) is the use of regression analysis of two time series for the purpose of examining a change in productivity levels. For example, the number of hours required to transport each prisoner by the police, in the pre-time period, was estimated with regression analysis (months being the units) in order to incorporate into the post-project time period any trends in the pre-project productivity level. The expected number of hours required to transport the actual number of post-project prisoners was then calculated by using the pre-project equation parameters (alpha and beta) on the actual number of prisoners transported. Since the total hours expected was more than the actual number required, the data indicate an increase in productivity.

A second analysis strategy of value in many evaluations is to identify the theory of the project in terms of how it is expected to have an impact

on the problem and then to test each of the linkages. In the DWI evaluation (p. 32 ff) the evaluator found an impact on a performance measure but the project had no effect on the intervening variables. This indicates either that the observed effect on the final performance measure is attributable to other factors or that the theory of the project was misspecified. The third technique of interest is the lagged regression analysis that was used to examine whether a crackdown on DWI would deter drunk driving and, therefore, reduce the number of accidents involving DWI (pp. 33 and 66-70). When the effect of an activity is not expected to occur simultaneously (e.g., within the time unit used in the analysis) a lagged regression such as that used here is appropriate.

E X C E R P T S

INTERIM PROGRAM EVALUATION:

D.W.I. IMPACT GRANT

PREPARED FOR THE WASHINGTON STATE
TRAFFIC SAFETY COMMISSION

STUART READIO
EVALUATION COORDINATOR
NORTHWEST REGIONAL COUNCIL

NOVEMBER 10, 1977

Summary

The data available at this early project date suggests that there has been little impact upon the number of accidents in rural Whatcom County in in which the driver involved was impaired by alcohol use. Individual program components, while maintaining a high degree of professionalism and, in fact, increasing overall productivity, have not in many instances exceeded statistical expectations in their activities.

The detention, apprehension and channeling of offenders into the judicial system as well as the optimization of the level of effort directed at this activity in order to install a high perception of risk of being apprehended for driving while drinking has certainly been implemented in a competent manner. Further, the judicial sanctioning of individual drivers to minimize recidivism appears to promise documented problem impact. It may well be a function of blunt measurement tools as well as the interim nature of this data that constrains our ability to discern a real program impact. This is certainly not to say that visible impacts are not expected by the second year of the grant period, at which time a more substantial data base should be available for evaluation purposes.

Briefly, this evaluation's major findings may be summarized:

- The Mobile Jail Van is indeed saving officer time and indirectly contributing to more effective law enforcement. In Blaine, officer time spent on transporting prisoners was down by 10%; the time processing DWI arrests was down 19.5% and the number of prisoner transports decreased by 11%.

- The DWI officer is exceeding expectations in terms of his participation in all Blaine Police activity. From November to August, Officer

Quaade contributed some 18.7% of the Department's total activity.

- Monthly DWI arrests in Blaine have increased sharply since the project began operation. With an F score of 4.06, DWI arrests established a new monthly trend, significantly greater than before the project began.

- The monthly number of physical control convictions have increased significantly during the project period, in the Blaine Municipal Court, with a F score of 4.07. We found convictions for this offense establishing a new and significantly greater monthly trend.

- State Patrol statistics reveal the DWI arrests increased significantly after the advent of the Mobile Jail Van. Patrol hours, the number of prisoner transports and the time spent by the Patrol on transporting prisoners also revealed positive progress toward reversing earlier unfavorable trends.

- The Alcohol evaluator made 75 alcohol evaluations during a 10 month period under the auspices of the grant. This is 18% of the total evaluations this person conducted.

- Probation clients previously recidivated at a rate of 26.5% in terms of prior DWI arrests. The overall recidivism rate for all activity is 44.6%

...

The Problem

In rural Whatcom County the number of accidents where the driver, upon investigation was found to be impaired by alcohol, had increased by 49% from 1974 to 1975. Further, the number of Canadians arrested for DWI

had increased signficantly (X^2 = 47.4, α = .004) from 1969 to 1973. From 1970 to 1976 DWI arrests by the Blaine Police Department had increased from 12 to 138, a change of 1050%. Traffic arrests had increased some 40% while police field contacts had increased 15%. Obviously the detection of DWI violators had increased dramatically in terms of the amounts of police time committed. Prior to the implementation of the DWI Impact Grant the Blaine Police averaged 11 DWI arrests per month, 1.5 physical control arrests and averaged 20 hours per DWI arrest. On top of all of this actual police patrol time, the meat of most departments' activity, had been decreasing from 6552 hours in 1971 to 3404 hours in 1976, a decrease in real patrol time of some 48%.

As incidious as these statistics may be, they have a more pervasive nature even yet. It was found that any increases in the number of DWI arrests made by the Blaine Police decreased the number of misdemeanant arrests, the police time spent on field contacts, the number of felony investigations the police made, the number of misdemeanant investigations the police made, and the number of juvenile contacts the Blaine Police made. Conversely, these same increases in DWI apprehensions increased police court appearances, the time the police spent in court and the number of prisoner transports by the Blaine Police to the Whatcom County jail.

None of the above-mentioned effects of increasing DWI apprehension can be construed as particularly positive in terms of police productivity and consequently in the deterrence of drinking and driving.

Program Methods

The DWI Impact Grant is a Traffic Safety Commission action program

designed to provide money, $100,687, to Whatcom County and the City of Blaine to fight the rising incidence of driving while intoxicated violations.

At the forefront of the local planning effort was the Director of the Whatcom County District Court Probation Department, Conrad Thompson, whose authored document, Drinking Drivers, identified specific crime problems particular to this area. The specific components of this project were then planned and implemented in Whatcom County to address this problem. The project level evaluation of all Impact Grant activities was funded contractually, $1,814, by the Impact Grant.

The City of Blaine and Whatcom County implemented a police patrol project as part of the overall program effort. This consisted of hiring and training a police officer, William Quaade, to deal specifically with driving while intoxicated violators. Officer Quaade was assigned to patrol the City of Blaine at those hours when the crime was most often thought to occur. The second aspect of this stepped-up increase in patrol activity was the purchase and outfitting of a step-van to act as a mobile jail unit. While Quaade's activity was aimed at increasing detections of the crime directly, the Mobile Jail Van sought to reduce police (in all of the smaller Whatcom County jurisdictions) down time and indirectly increase crime detections. It was also assumed that the Mobile Jail Van would act as a deterrent to all types of criminal activity.

The second component of this grant was to increase prosecutorial services to the city of Blaine, again directed at increasing the thoroughness and efficiency with which crime specific activities are handled. The grant made monies available to increase the time an attorney could act on

the city's behalf in the processing of apprehended DWI violators. Further, to insure the proper adjudication of these court cases the grant made available monies for increased judicial services to the City of Blaine. These monies were used to allow the City to contract with Whatcom County for the use of its circuit-riding judges.

At the end of the Impact Grant's spectrum were the important diagnostic and rehabilitative functions performed by an alcohol counselor and probation officer. The alcohol evaluator is a staff member of the Whatcom County Alcohol and Information and Referral Center, while the Probation Officer is a member of the Whatcom County District Court Probation Department. It is the responsibility of the Alcohol Evaluator to diagnose potential alcohol problems amongst all drinking drivers convicted in the Blaine Municipal Court. The probation officer, Alex Whitehouse, has the responsibility of gathering sentencing information, assuring that referrals are made for alcohol evaluations, and insuring accountability for programs required of the offender by the Court. These functions apply to both Canadian and American citizens.

Supplemental to all of these components, including the evaluation, was the provision of clerical services to the City of Blaine's Police Department and Municipal Court. Willie Kilmer handles these duties as police secretary and court clerk with sterling acumen. Her orderliness added immeasurably to the coherence of this evaluation.

It should be mentioned that in no case was new staff hired to specifically attend to project endeavors. In most cases, existing staff were utilized to fill these new positions and often the grant actually merely made monies available for the work to be done. In terms of program

efficiency, this proved to be quite good as DWI detection, adjudication and rehabilitation efforts were manned by experienced staff. In terms of this research, it made the work somewhat confusing.

This is certainly a far-reaching project, whose impacts were direct and indirect. I have focused on the most obvious components and paid, perhaps, too little attention to some of the less obvious aspects. The second year of evaluation activity will certainly attempt to correct any such oversights.

Time-Series Statistics and Related Research

Much of the data in this document has been examined by means of the interrupted time-series design and several accompanying statistical tests. The use of interrupted time-series analyses presupposes, among other things, the existence of a valid meausre of the effect variable (this DWI counter-measures project) for an extended period preceding and following the change in the causal variable. Within this evaluation I have gone back to 1973 to collect monthly activity data from both the Blaine Police Department and the Municipal Court. This monthly data was plotted up until November of 1976 as the "pre" period. Data subsequent to this period was used as the "post" period and hypothetically supposed to show the impact of the project. Actually, in all of these time series calculations it was hypothesized that the pre trend would be equal to the post trend; in other words, the monthly data would establish a direction and remain unchanged despite intervention. To test for change due the DWI counter measures project I have used tests of significance, the Mood, and the Walker-Lev l. The Mood test the t-distribution while the Walker-Lev test uses the F-distribution of probabilities.

The single Mood examines whether there was any immediate change in the trend established by the pre data. It does this by predicting the first post data point and then comparing this predicted value with the first actual data point. A significant difference may mean that there was an "at-once" effect upon the data brought about by the introduction of counter-measures.

The Walker-Lev 1 test is a test of the hypothesis of a common slope for the pre change and post change periods. In other words, the hypothesis states that all post change data points fit the slope of the pre change data. If an impact is made upon the variables tested then the slopes should not be common.

Since the DWI counter-measures project intends to stimulate a new trend in the data, the Walker-Lev 1 test will be emphasized here.

In conclusion, we have a test, the Mood, that will show us whether the project had made an immediate impact upon the project.

The Walker-Lev 1 will reveal whether the project has created a new trend in the data, a trend we suppose will include greater post change values and yet hypothesize will not.

To analyze the jail van data I have collected activity statistics from the Blaine Police Department and constructed regression equations which allowed me to predict post change monthly scores. These scores were then compared with actual monthly statistics...

For both the DWI officer and the mobile jail van descriptive statistics are breifly examined. These data were collected by the program people in the court of their activities and later aggregated by the project director. Though they do not include a twelve month period of data, they are

a sufficient sample and may be considered representative of the whole population.

The rehabilitation aspect of this project was examined by means of substantial data collection activities and aimed at establishing a baseline recidivism rate that a thorough follow-up might be made on clients during the second year of the project. Demographic and legal history variables were also collected on clients to act as predictors of future criminality. The process at work on clients was also investigated to help explain the outcome. Unfortunately scant information was maintained on and by the alcohol evaluator, though this problem is modified somewhat by an overlap of data with the Probation Department.

Privacy and security agreements have been signed with the Probation Department assuring access as well as protecting client security.

Assessing whether overall problem impact had occurred was constrained by a definite lack of data. I was able to collect impaired driver accident data only up until July of 1977. This of course means that our post change period was rather lacking in data points. An analysis at the end of the second year showed alleviate this problem. I selected rural Whatcom County as the geographic location that I might forego any extraneous, policy-oriented influences on city data. An examination of 1976 impaired driver accident data for the city of Blaine revealed zero accidents and zero fatalities. To analyze what data I did have, I lagged the correlation between accidents, arrests by the Blaine Police, convictions of DWI violators by the Municipal Court as well as probation referrals to alcohol treatment of DWI violators. This lagging of the data allowed me to test the deterrent effect of project emphasis on these three major variables.

With the resulting equations using pre change monthly data, I then predicted monthly totals and compared them with actual monthly scores.

...

Time-Savings of Mobile Van

One assumption inherent in the funding of the mobile jail van was that it would save officer time at booking, testing and transporting suspects. Such time savings should, of course, manifest itself in terms of increased police productivity (more apprehensions). Rather than doing so directly, several intervening variables should have been impacted, thus allowing increased crime detection productivity. I have isolated four distinct intervening variables that should have changed with the advent of the mobile jail. They are: a) the number of prisoner transports; b) the police time spent on prisoner transport; c) the amount of police patrol time; d) the time spent processing DWI arrests. Variables a, b and D should decrease while variable c should increase. Instead of researching what are probably non-existent statistics from all of the smaller Whatcom County law enforcement agencies, I have collected these statistics from the Blaine Police Department. I believe this to be a logical sample frome as the grant did target Blaine and we have seen Blaine make copious use of the mobile jail. Further, the exemplary nature of the majority of the Blaine Police data makes the research less prone to invalidity.

To test the time savings inherent in the mobile jail van I correlated patrol time with the number of DWI arrests, on a monthly basis. After calculating the slope and y-interrupt I constructed a regression formula, $\hat{y} = 5.73 \text{ x } +257.48$ which allowed me to predict for the during-program months how much patrol time should have been utilized according to the number of

arrests. I expected 3400 hours of patrol time while the actual amount was 11% less, or 3051 hours. In other words, it does not appear as if the Blaine Police Department's patrol time has changed significantly, especially in light of the fact we should have expected an increase in patrol time if the van had made an impact.

The same technique was utilized to predict the number of prisoner transports. In this case the predictor variable was the amount of time spent by the police on transporting prisoners. The regression equation in this case was $\hat{y} = .51 x + 3.43$. I expected the Blaine Police to make 115 prisoner transports while in fact they only made 102, a difference of 11%. In this case then there has been a decrease in a variable that may certainly have been brought about by the mobile jail van.

The amounts of police time spent in transporting prisoners was also regressed, in this case, against the number of prisoner transports. The regression equation was $\hat{y} = .91 x + 8.19$. According to this equation I expected 175 hours to be spent by the Blaine Police transporting prisoners while I found that they actually spent 158 hours, a decrease of 10%. Again this is a promising statistic as regards the impact, in Blaine at any rate, of the mobile jail van.

One further calculation concerning the van's impact upon the City of Blaine was investigated and this involved the pre and post van police time spent on DWI arrests. I have used the same technique as above. I correlated the number of DWI arrests with the police hours spent on these arrests for 24 months prior to the advent of the van. This correlation revealed the forumla for predicing time spent of $\hat{y} = 1.73 x + .55$. Then, by substituting the post van number of DWI arrests I <u>predicted</u> the hours the Blaine Police <u>should</u> have spent processing DWI arrests. The total was

225-1/4 hours. By reviewing monthly police activity reports for this same period, I found that the police actually spent 181.25 hours, a substantial decrease of 19.5%. This may be a function of an increased proficiency on the part of individual officers in handling cases though I suspect, as with the former variables, a minifestation of the impact, of the mobile jail van on the Police in Blaine.

In conclusion, we have seen three important variables, the number of prisoner transports, the time spent on prisoner transports and the time spent on DWI arrests aggregate during the project period to a total less than expected. The nature of these variables are such that they should have been amendable to change with the advent of the mobile jail van. Police patrol time, on the other hand, continued its downward trend and though it should have increased, it did not. These mixed results may well be the product of incomplete data but they do appear to answer the immediate question as to whether there is any manifest van impact.

It can only be hypothesized that similar results may be found in other county law enforcement statistics. If they are similar; then we may say that the jail van output is up to expectations.

...

Van Impact Upon Washington State Patrol Statistics

Certain activity statistics from the Bellingham Detachment of the Washington State Patrol were also utilized to measure the impact of the Mobile Jail Van. The monthly statistics, patrol hours, number of prisoner transprots, time spent by WSP officers on prisoner transports and DWI arrests were selected for their perceived amendability to change induced by the Mobile Jail Van. These figures were analyzed by means of the Mood and

Walker-Lev 1 test used and described earlier. The findings are summarized below.

WSP Activity and Time-Series Analysis

Activity	Test	Score	Degrees of Freedom	Significance
Monthly Patrol Hours	Mood	t= .37	33	NS
	Walker-Lev 1	F= .08	1,41	NS
No. of Prisoner Transports	Mood	t= -.34	33	NS
	Walker-Lev 1	F= 1.01	1,41	NS
Time on Prisoner Transports	Mood	t= .49	33	NS
	Walker-Lev 1	F= .23	1,41	NS
DWI Arrests	Mood	t= .19	33	NS
	Walker-Lev 1	F= 6.04	1,41	p= .05

The data seem to indicate that in no case was there an immediate effect or change in State Patrol activity. In only one case was there an apparent long term change in the monthly data and that was the number of DWI arrests. Unfortunately, this data point is perhaps the least amendable to being changed by the advent of the Mobile Jail Van. The assumption is that greater efficiency in booking and transporting suspects would facilitate and increase the time available to the State Patrol for detecting offenders. This is plausible, however, there is the intervening step of increasing patrol time and subsequently, increasing the "at-risk" period for drinking drivers. No significant change in the amounts of patrol time available does seem to cloud the issue in terms of the van contributing directly to greater WSP drinking driver apprehensions. There are too many possible extraneous influences upon DWI arrests to allow us to say unequivocally that the van brought about this change.

...

Problem Impact

As I mentioned earlier, the lack of data has constrained the ability of this research to discern an overall project impact on the problem of drinking and driving violators and the accidents resulting therefrom in rural Whatcom County.

To measure the effect of increased police apprehensions, DWI court case convictions and probation referrals to alcohol treatment, I have correlated these monthly statistics with the monthly number of impaired driver accidents in rural Whatcom County. To better determine the deterrent effect of these former variables, I lagged the accident data one month. In other words, July of 1976 arrest data would be correlated with August of 1976 accident data; and then August of 1976 arrest data would be correlated with September of 1976 accident data and September of 1976 arrest data would be correlated with October of 1976 accident data, on through until I run out of monthly data points. This technique should reveal whether increasing DWI apprehensions, convictions and the referral of alcohol offenders to treatment acted in a manner sufficient to deter drinking drivers and consequently effect impaired driver accidents. Further, these calculations allow us to predict the monthly number of accidents which can then be used in a comparison with actual monthly statistics.

This r^2 between accidents and arrests was .01 which is extremely low. The regression forumla used for predicting accidents turned out to be $\hat{y} = .10 \times +11.64$. Utilizing this formula and using actual arrests to predict accidents, we arrive at totals for 89 for expected accidents while the actual total was 98.

The lagged correlation between DWI convictions and impaired driver accidents was slightly higher (r^2 = .03). The regression equation of y = 4.94 x +8.8 revealed a predicted accident value of 89 also.

Though hampered by even less data (perhaps a partial explanation) I found that the lagged correlation between accidents and probation-directed alcohol referrals was quite a bit greater (r^2 = .11) and that the differences between expected value for accidents was less than the actual total. Again, this should be examined with some caution as I am using for this regression 2 months less data than in the other correlations. The actual worth of this final statistic will be known later when more data is available. It does at this stage seem somewhat more promising than the others. This is probably because the scope of probation referrals is greater than Blaine Police arrests and Municipal Court's convictions. In none of the correlations do I have much faith in the predictability of the regression equations due to the rather low r^2.

...

SECTION 5

PROTECTING THE CONFIDENTIALITY AND PRIVACY OF DATA

Overview

This section contains a table showing the different forms or assurances that are to be used and when each is needed. The full text of LEAA's most recent explanation of their requirements is included and sample of forms actually used in an evaluation are shown.

SECTION 5A

PROTECTING THE CONFIDENTIALITY AND PRIVACY OF DATA

In examining the issues about protection of confidentiality and privacy of research or statistical data, it is useful to make a distinction among four elements:

1. What is required and what is permitted by the LEAA regulations.

2. What is required and what is permitted by the state law and administrative code (WACs).

3. What should be done by any agency involved in the collection, transfer, or use of statistical or research data in order to protect the agency from possible litigation by individuals described by the data.

4. What should be done, from an ethical rather than legal point of view, in weighing the value of the research or evaluation against the possible risks to individuals involved as "subjects" in the evaluation and/or from whom the data are collected. These risks include possible invasion of privacy, accidental or intentional revelation of the confidential information about a particular individual to another person without the informed consent of the first individual, and any other negative consequences that might occur for the individual if the evaluation is conducted and/or the data are obtained, transferred, or used for research and statistical purposes.

In a general sense, it is accurate to say that officially promulgated statements on requirements are not as stringent as what would be needed to fully protect an agency or an evaluator against potential litigation. Furthermore, the procedures needed to protect an agency or evaluator against litigation may not be as stringent

as those which would meet all of the ethical standards that one might propose to be needed to fully protect human subjects.

This section of the Handbook begins with a summary table showing the names of the written forms that may be needed, summary information about them, and the location (in the Handbook) where a discussion and sample of each form can be found. The summary table is followed by a discussion of informed consent and protection of human subjects.

A full copy of the LEAA regulations (which are quite clear and self-explanatory) is included following the text of this section. In the final part of this section are the forms used to obtain names of juveniles from an agency for the purpose of contacting and interviewing the youths. Similar procedures could be used, with adaptations to meet specific needs of the situation, by evaluators attempting to gain access to similar types of sensitive information.

AGREEMENTS AND ASSURANCES FOR COLLECTION, TRANSFER, AND USE OF DATA ABOUT INDIVIDUALS TO BE USED BY EVALUATORS FOR RESEARCH OR STATISTICAL PURPOSES

NAME OF FORM OR ASSURANCE	PARTIES TO THE AGREEMENT	PREPARED BY	WHEN NEEDED	PURPOSES	EXAMPLES & DISCUSSION OF ELEMENTS IN FORM
1. Transfer Agreement	(a) agency that collects information & evaluator/s to whom it is being transferred; (b) evaluator's agency & any other agency or individual (other than employees or subcontractors) to whom it is being transferred.	Recipient or transferrer of data	(a) when data are identifiable or potentially identifiable to individuals; (b) when names are released.	Specific agreements & conditions to protect confidentiality & privacy of data.	Handbook, Section 5D; LEAA Regulations, pp. 11-12, 25-26, 33
2. Privacy Certificate	Issued by evaluator or evaluator's agency to agency that is transferring the data.	Recipient of data	Same as above.	Assurance by recipient that data will be kept confidential & information on how this will be done.	Handbook, Section 5D; LEAA Regulations, pp. 21-24, 29
3. Employee Agreement	Employees of the evaluator's agency who will have access to the data.	Evaluator	Same as above.	Same as above.	Handbook, Section 5D
4. Sub-contractor Agreement or Provision in Sub-contract	Evaluator and subcontractor	Evaluator or subcontractor.	Same as above.	Same as above.	Handbook, Section 5D

[CONTINUED ON NEXT PAGE]

AGREEMENTS AND ASSURANCES (continued)

NAME OF FORM OR ASSURANCE	PARTIES TO THE AGREEMENT	PREPARED BY	WHEN NEEDED	PURPOSES	EXAMPLES & DISCUSSION OF ELEMENTS IN FORM
5. Informed Consent	Subjects from whom data are being collected.	Evaluator or person wanting the data.	When collecting data directly from individuals for evaluation, research or statistical purposes.	To protect human subjects.	Handbook, Section 5D; LEAA Regulations, pp. 14-15
6. Protection of Human Subjects Review Form	Outside committee review of procedures.	Evaluator prepares information; committee approves or disapproves the procedure.	Required by DHEW & some other federal agencies; not required by LEAA but useful to insure that ethical & legal requirements are met when human subjects are involved.	To protect human subjects.	Handbook, Section 5E

SECTION 5B

DISCUSSION OF INFORMED CONSENT AND HUMAN SUBJECTS REVIEW PROCEDURES

Risks to the individual and the possible invasion of an individual's privacy are substantially increased when the evaluator contacts the individual to collect data directly. Evaluators almost always must be concerned with the risks to human subjects when the project, for purposes of research or evaluation, manipulates human beings in a manner that would not otherwise have occurred.

The issues to be covered in the subsequent discussion are: (1) What procedures should the evaluator follow in obtaining the names of persons to be contacted and interviewed directly? (2) What constitutes "informed consent"? (3) What is involved in the use of a human subjects review committee?

Obtaining the Names of Persons to be Contacted

An evaluator can find a catch-22 situation in which s/he must contact individuals to obtain informed consent, but is unable to get the names of project clients in order to contact them to obtain the consent. The LEAA regulations do not require any agency to release names to an evaluator so that the latter can contact the individuals to obtain informed consent, but the regulations do not in any way prohibit this.

One of two procedures generally would be used to obtain names of persons who are to be re-contacted by the evaluator. The first would be for the project to contact the individuals and administer the informed consent procedure on behalf of the evaluator. Names of persons who agreed to be contacted would then be released to the evaluator upon the

completion of a transfer agreement and privacy certificate. The consent of clients to be contacted by evaluators could be obtained by project personnel at the time the individual is a client of the project. The transfer agreement and/or privacy certificate should include a specification of any risks to the client that might be involved in the evaluator's effort to contact them and the procedures to be used for minimizing the risk. If the project assumes responsibility for contacting the clients, it is presumed that the project only seeks consent for the evaluator to be given the names so that s/he can contact the individual and seek informed consent. (The project could, of course, seek consent for the individual's participation in the interview or data collection procedure, but the evaluator probably would prefer to do this.)

A second procedure would be for the evaluator and agency to reach agreement on release of the names to the evaluator and the procedures to be used by the evaluator for contacting individuals and obtaining informed consent (details should be included in the transfer agreement and/or privacy certificate). The agency could then notify the client that s/he is to be contacted by the evaluator, but does not have to consent to be included in the interview or other data collection procedure. The evaluator would contact the client and seek informed consent. This second procedure, again, is neither required nor prohibited by existing regulations. If it is used, both the project and the evaluator should be aware of the potential risks in even contacting the individual. For example, efforts to recontact a rape victim could result in persons currently unaware that the rape occurred (parents, husband, or children, for example) becoming aware of it.

The agency that transfers names to evaluators should balance the benefits to be gained by the evaluation or research against the risks to the individuals, including but not limited to the invasion of privacy and violation of confidentiality. The privacy certificate that is filed along with the transfer agreement should include a discussion of the benefits, the risks, and the procedures used to minimize the risks.

Informed Consent

Before the evaluator collects information directly from individuals s/he should obtain written informed consent from the individual. The person from whom data are to be collected should be accurately informed about the research, fully informed of risks (if any) involved in agreeing to participate, and informed as to whether there are any penalties for withdrawing after initially agreeing to participate. The individual should also be informed as to what will be done with the data (e.g., used for research purposes), whether the data are confidential, who will use the data, and what will be done with it when the evaluation is complete. The evaluator, of course, can include information concerning the benefit of the research to the individual, community, or society in order to encourage participation. The purpose of informed consent is to insure that persons do not enter into an evaluation study without adequate information about the purposes and risks of the study. In addition, the individual must voluntarily choose to participate.

Human Subjects Review Committee

Some federal agencies require that a committee which includes at least some persons outside of the agency conducting the research review

the protection of human subjects procedures to be used in the evaluation. The general procedure is that the evaluator completes a "human subjects review form" (a copy is included in the Handbook) and the committee either approves or disapproves of the procedures. If the committee disapproves, the evaluator would need to alter the procedures and re-submit the request. Some organizations, including most universities and many non-profits, utilize a human subjects review committee (even when it is not required) in order to further insure that ethical and legal requirements are met.

SECTION 5C

LAW ENFORCEMENT ASSISTANCE ADMINISTRATION REGULATIONS: "CONFIDENTIALITY OF RESEARCH & STATISTICAL DATA"

Abstract

The full text of LEAA's most recent explanation of the agreements, assurances, and procedures for protecting human subjects and data about them is included. This is a well written presentation that clarifies, interprets, and explains the requirements originally published in the Federal Register of December 15, 1976 (vol. 41, no. 242, pp. 54846-54848).

Confidentiality of Research and Statistical Data

U.S. Department of Justice
Law Enforcement Assistance Administration
National Criminal Justice Information and Statistics Service

James M. H. Gregg
Acting Administrator

Harry Bratt
Assistant Administrator
National Criminal Justice Information and Statistics Service

Carol G. Kaplan, Director
Privacy and Security Staff

Law Enforcement Assistance Administration
U.S. Department of Justice

Preface

This document was prepared by the Privacy and Security Staff, National Criminal Justice Information and Statistics Service, in conjunction with the LEAA Office of General Counsel, to explain and discuss the requirements of the LEAA regulations governing confidentiality of research and statistical data (28 CFR Part 22). It is hoped that the document will clarify some of the requirements and objectives of the regulations and will serve as a guide to persons conducting research and statistical activities pursuant to LEAA-funded projects. Of equal importance, it is hoped that the document will provide potential project subjects with an easily understood statement of the scope and protections of the regulations.

Background

The Statute

The LEAA regulations on confidentiality of research and statistical data, which are contained in 28 CFR Part 22, implement Section 524(a) of the Omnibus Crime Control and Safe Streets Act of 1968, as amended. Section 524(a) provides that:

> Except as provided by Federal law other than this title, no officer or employee of the Federal Government, nor any recipient of assistance under the provisions of this title, shall use or reveal any research or statistical information furnished under this title by any person and identifiable to any specific private person for any purpose other than the purpose for which it was obtained in accordance with the title. Copies of such information shall be immune from legal process, and shall not, without the consent of the person furnishing such information, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceedings.

5-14

The section was enacted as part of the 1973 amendment to the Omnibus Crime Control and Safe Streets Act.

The Regulations

In recognition of the significance of the issues involved, draft regulations implementing the act were initially published in the *Federal Register* on September 24, 1975. Public hearings were conducted on October 11, 1975; written comments were also received from interested groups. Subsequently, on January 8, 1976, an ad hoc panel of interested persons was convened to discuss proposed revisions to the draft regulations. The panel included representatives of the criminal justice, academic, and research communities, as well as representatives of other interested Federal agencies. On December 15, 1976, final regulations were promulgated in the *Federal Register*.

Objectives

The objectives of these regulations are:

(1) to ensure the confidentiality of identifiable data collected for a research/statistical purpose;

(2) to upgrade the validity of research findings (by minimizing subject concern over subsequent use of personal information); and

(3) to clarify researchers' obligations, responsibilities, and protections with respect to use and revelation of identifiable research/statistical data.

Impact

Summary of Requirements

In summary, the regulations provide:

- o that identifiable research/statistical data may only be used (without consent of the individual) for research or statistical purposes;
- o that data may only be transferred in identifiable form pursuant to a transfer agreement ensuring recipient compliance with confidentiality limitations;
- o that, except in noted circumstances, subjects must be advised that data will only be used for research or statistical purposes;
- o that, upon completion of a project, identifiers must be destroyed or otherwise separated from data and permanently secured; and
- o that copies of identifiable data are immune from administrative or judicial process.

Protection to Subjects

The regulations are intended to ensure that information provided for research/statistical use:

- o is not transferred or revealed in identifiable form for any purpose other than additional research or statistical activity (without prior consent of the individual);
- o is not used for purposes other than research/statistical activity (without prior consent of the individual);
- o is not included in identifiable form in reports or publications (without prior consent of the individual); and
- o is maintained under physically and administratively secure conditions to protect against unintentional revelation of identifiable data.

The immunity sections of the regulations ensure that copies of information identifiable to a private person collected for research/statistical purposes cannot be subpoenaed or otherwise legally compelled to be produced in a judicial or administrative proceeding.

Key Concepts

The major areas to be considered in understanding the requirements of the regulations are grouped under the following general subject headings:

Applicability

Information Protected

Information Not Protected

Authorized Uses and Transfers of Data

Subject Notification Requirements

Final Disposition of Data

Security of Data

Immunity

Applicability

The regulations *apply* to all projects that:

- were funded with LEAA funds awarded subsequent to January 14, 1977; and
- involve the collection of information identifiable to a private person for research/statistical purposes.

Coverage extends to research or evaluative "components" of LEAA-funded "action or delivery" projects in cases where identifiable data is collected in the evaluation or research component.

The regulations *do not apply* to:

- projects in which data are collected for a research or statistical purpose in unidentifiable form;
- "action" projects in which identifiable data are collected for administrative or operational uses;
- projects where funds are used for development of data-collection capability--rather than collection of data.

Statutory requirements, including the immunity provisions, apply to all personally identifiable research/statistical data collected in LEAA projects funded after July 1, 1973.

On October 3, 1977, Public Law 95-115 incorporated Section 524(a) into the Juvenile Justice and Delinquency Prevention Act of 1974. Research and statistical projects funded under this Act are now fully subject to the Section 524(a) confidentiality provisions.

Information Protected

The statutory language on which the regulations are based provides that the protections of the statute apply to *research and statistical information* that is *identifiable* to any *specific private person*. The protections apply regardless of the nature, subject matter, or "privacy implications" of the information.

Research and statistical information is defined by the regulations as: "information obtained for a research or statistical purpose in a project (or project component) whose objective is to test, measure, evaluate, or otherwise increase knowledge in a given substantive area."

Under this definition:

- o Identifiable information obtained for administrative or "housekeeping" purposes is *not* protected--even where obtained in connection with conduct of a research/statistical project.
- o Identifiable information obtained for research/statistical purposes in a component of an otherwise "action/delivery" program *is* protected under the regulations.

Private person is defined by the regulations as including corporations or nongovernmental organizations, as well as individual persons. The term includes persons (such as law enforcement officials) operating in an official capacity, but excludes governmental agencies.

Under this definition:

- o Information identifiable to a law enforcement officer whose activities were the subject of a research or statistical effort *would* be confidential under the regulations.
- o Information identifiable to a particular police department *would not* be confidential under the regulations.

Identifiable information is defined by the Regulations as information which may "reasonably" be identified to a private person. The term is to be construed on the basis of factors such as:

- o the size of the statistical universe;
- o the availability of public records that could be combined with research data to reveal an individual's identity;
- o the uniqueness of certain attributes of subjects; or
- o inclusion of a variety of demographic characteristics on the subjects.

Information Not Protected

The following categories of information are not covered by the regulations and, as such, may be released in identifiable form for any purpose:

- o information obtained from records designated under State or Federal statute as "public" (exempted to preclude conflict with State open-record policies and "sunshine" legislation);
- o information regarding future criminal conduct (exempted to preclude conflict with Federal and/or State law);
- o information gathered for intelligence or law enforcement purposes (exempted to ensure that "intelligence" data are not included as "research and statistical information").

In addition, where identifiable data is obtained from *non-public* records for research/statistical purposes, the regulations apply to the extracted research/statistical data *only*. This exemption is specifically stated to preclude law enforcement concern over the possible extension of applicability of confidentiality regulations to administrative or criminal history record systems from which data is released.

It should be noted that the regulations do not require disclosure of the information described above. Accordingly, researchers may *voluntarily* withhold disclosure of such information--recognizing, however, that where such information is sought pursuant to a subpoena, information would not be protected by the immunity provision of the act and regulations.

Authorized Uses and Transfers of Data

The regulations provide that identifiable research/statistical data may only be used or revealed--on a need-to-know basis--as follows:

- o for other research or statistical purposes;
- o for any purposes authorized by the individual subject;
- o to employees of the recipient of assistance;
- o to subcontractors (provided subcontracts contain provisions to ensure security, confidentiality, and return of identifiable data); and
- o to LEAA--for limited statutory reporting and auditing purposes.

The regulations do not:

- o limit eligible recipients of data for research/statistical purposes or require that researchers be certified or licensed for the purpose of obtaining data;
- o *require* that data--in identifiable or nonidentifiable form--be transferred for secondary use;
- o require that LEAA approval be obtained prior to transfer of data.

Transfer for Research/Statistical Purposes

Information may be transferred or revealed in identifiable form for any research or statistical purpose. To ensure confidentiality of the information, however, the regulations require that:

- Information may only be transferred in identifiable form on a *need-to-know basis* (thus requiring that identifiers be stripped where transferred data can be utilized without identifiers).

- Information may only be *re*transferred where data are included in the recipient's data base and transfer is approved by the original transferor of data.

- Information must be returned upon conclusion of the project for which data is transferred unless alternative arrangements, consistent with the regulations, are agreed upon.

- Information may only be transferred pursuant to a *transfer agreement* binding the recipient of data to the restrictions of the regulations.

- Information may only be transferred upon a finding by the transferor that:

  . the proposed research use will not cause social or economic harm to the individuals identified in the data to be transferred;

  . the proposed project will be designed to ensure confidentiality; and

  . adequate administrative and physical security of data will be maintained by the recipient of the data.

Transfer with Consent of the Individual

The regulations provide that identifiable information may be revealed or transferred for nonresearch or statistical purposes where prior consent has been obtained from the individual to whom the information relates.

- Issues relating to consent (e.g., competence of consenting individual, persons authorized to consent for minors, etc.) will be determined pursuant to applicable State law.

- Subject consent may *generally* be obtained at any time prior to release or use for nonresearch/statistical purposes (including at the time of data collection).

- Where the data are sought for use in a judicial or administrative proceeding, however, written consent must be obtained at the time that the data are sought for use in such proceedings.

- Although not specifically stated in the regulations, it is recommended that all consent be obtained in written form and that copies of the consent be retained by both the persons releasing and receiving the data.

Subject Notification Requirements

The regulations distinguish among situations in which:

- o data are obtained *directly* from the subject through questionnaire or other direct inquiry;
- o data are developed through *observation* of subject activity; or
- o data are derived from *existing* records.

Specifically, the regulations require that:

Direct inquiry: Where data are obtained through direct subject inquiry, subjects must be advised, either orally or in writing, that information will, in the absence of alternative notification and consent, be used for research or statistical purposes only and that participation is (or is not) voluntary.

Subject observation: Where data are obtained through direct subject observation, subjects must be advised of the above-noted facts as well as the types of information to be collected.

Existing records: Where data are obtained from existing records, no notification of subjects is required. (Data obtained in this manner are, however, in all other respects covered by the provisions of the regulations.)

Waiver of Notification

Subject notification requirements may be waived where information is to be obtained through direct observation and, in the view of the researcher, notification would preclude or seriously impede conduct of the project. In such cases a justification for the waiver must be included as part of the privacy certificate.

"Unique" Subjects

Where data are obtained directly from a subject, the subject must be advised if it appears--by virtue of sample size or subject uniqueness--that identity cannot be reasonably concealed. In such cases, agreement to participate in the study is deemed to constitute consent to revelation of data in potentially identifiable form in research/statistical products of the project. Agreement to participate does not, however, without additional specific consent, authorize disclosure of the data for nonresearch/statistical uses.

Security of Data

The regulations require that physical and administrative security of identifiable research/statistical data be ensured by the original researcher and by all subsequent recipients of data.

To accomplish this objective, the researcher must:

- o notify all staff (paid or volunteer) of the requirements of the regulations and obtain written agreement therewith from all employees;
- o limit staff access to identifiable data on a "need-to-know" basis;
- o maintain data under physical conditions designed to preclude intentional or accidental access to data by nonauthorized individuals.
- o maintain a log indicating all transfers of information in identifiable form. (The log should indicate the name of the individual to whom the information was released, the individual's organization, the date of dissemination, identification of records released, and the purpose for which the transfer was made.)

To ensure proper administrative security, it is also recommended that a list of all individuals (including employees) *authorized* to have direct access to the identifiable data base be developed and that a record of *actual access* to identifiable information by these authorized users be maintained.

Computer Storage

Where identifiable data are to be maintained in a computer, the researcher must obtain written assurances that adequate hardware and software and administrative procedures will be utilized to:

- o ensure technical security of data;
- o preclude unauthorized access to identifiable data; and
- o prevent unauthorized linkage of data.

There is no requirement that data be stored in a "dedicated" system or that data be entered in the computer in nonidentifiable form.

Immunity

The regulations (in Sec 22.28) follow the language of the act and provide that:

> "Copies of research or statistical information identifiable to a private person shall be immune from legal process and shall only be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding with the written consent of the individual to whom the data pertains."

In interpreting the immunity provisions, the following should be noted:

- o The statute provides for an automatic immunity, requiring no action by the researcher or LEAA.
- o Immunity applies to "copies" of data--and accordingly, would not apply to researcher recollections of information, nonrecorded impressions of subject response, etc.
- o Immunity applies regardless of whether or not subjects are notified of project participation or of the immunity protections.

- o Immunity is only applicable to "administrative" and "judicial" proceedings and accordingly would not protect against release of data in legislative proceedings.
- o Immunity is based on the Federal statute and not merely on the language of the regulations or a grant condition.
- o Immunity applies to research/statistical information collected after August 1973 (LEAA-funded) and October 3, 1977 (JJDP-funded)--regardless of whether or not the project received additional funding after that date.
- o Immunity applies to Federal and State court or administrative proceedings.

Final Disposition of Data

Upon termination of a project in which identifiable data were collected, the regulations provide two options with respect to final disposition of data.

Specifically:

- Data, or all identifying portions thereof, may be destroyed.
- Identifiers may be stripped from data and a name-index retained under separate and secure conditions.

Removal of identifiers and maintenance of a separate name-index code will permit subsequent longitudinal studies and/or reanalysis of data. Where the name-index procedure is to be followed, a description of the separate maintenance conditions for the name-index must be included in the privacy certificate.

In planning for the final disposition of data, researchers should be aware that Federal or State requirements may preclude the destruction of records for a specific period of years from the completion of the project. In such cases the name-stripping process should be utilized during this period.

Disposition of Transferred Data

Identifiable data transferred to a subsequent researcher pursuant to a transfer agreement must be returned to the original researcher at project conclusion unless alternative arrangements are agreed upon. Such arrangements must include, *as a minimum*, the maintenance of a separate and secure name-index code.

Procedural Requirements

PRIVACY CERTIFICATE TRANSFER AGREEMENT

The procedural mechanisms through which the regulations will be implemented are the *privacy certificate* and the *transfer agreement*.

Copies of a sample privacy certificate and sample transfer agreement are included at the end of this document. These forms are *samples only*. Alternative forms may also be used so long as they contain the necessary assurances.

The Privacy Certificate

The regulations require that a privacy certificate be submitted as part of any application for a project in which data identifiable to a private person will be collected for research or statistical purposes. A certificate must, therefore, be submitted in connection with research/statistical projects and with those "action" projects which include an evaluation component involving the collection of data identifiable to a private person. A certificate would not be required in projects in which data is to be collected in nonidentifiable, statistical form only.

Contents of Privacy Certificate

The privacy certificate should contain assurances that:

- o Limitations on use/revelation/transfer of identifiable data will be maintained.
- o Adequate administrative and physical security procedures will be undertaken.
- o The project and any project reports will be designed to ensure confidentiality of data.
- o Appropriate subject notification procedures will be followed.
- o Dissemination log procedures will be followed to control release of identifiable data.

To support these assurances, the privacy certificate should briefly describe:

- o procedures to ensure confidentiality of data;
- o procedures to ensure physical/administrative security of data;
- o procedures for subject notification and/or justification for waiver thereof (pursuant to Sec 22.27(c) of the regulations); and
- o procedures for final disposition of data (including security arrangements where separate name-index codes will be maintained).

The certificate should also include the name and title of:

- o the individual to be charged with primary responsibility for ensuring compliance with the regulations (generally, the project director);
- o the individual authorized to approve transfers of data (and any institutional limitations associated with data transfer); and
- o the individual authorized to determine final disposition procedures for data developed in the project.

Where relevant, a copy of consent forms should also be attached to the certificate.

Submission of Privacy Certificate

A privacy certificate should be submitted as part of any application for a project to be funded under the Omnibus Crime Control and Safe Streets Act in which research/statistical data identifiable to a private person is to be collected.

Where applicants do not initially anticipate that research/statistical data will be collected in identifiable form, a privacy certificate should be submitted and approved at such time as funds are, in fact, to be expended for collection of identifiable research/statistical data. In such cases, a special condition may be included in the grant providing as follows:

> "Where a privacy certificate is not initially submitted, such a certificate must be submitted and approved prior to the expenditure of LEAA funds for collection of identifiable research/statistical data."

Privacy certificates may be amended at any time, subject to approval by the appropriate reviewing official or board.

The Transfer Agreement

The transfer agreement is intended to ensure the confidentiality of identifiable information which is transferred from the original LEAA supported researcher to a subsequent researcher. Although a transfer agreement is required in connection with each transfer of data, successive transfers of data to the same recipient may be handled through amendments to an original agreement.

Contents of Transfer Agreement

The information and assurances to be included in the transfer agreement are indicated in the sample transfer agreement which is included at the end of this document. Where institutional regulations require that additional assurances be obtained in connection with transfer of data, such assurances must also be addressed prior to transfer of data.

The transfer agreement should be signed by the individual authorized to transfer identifiable data protected under the regulations, as indicated in the original privacy certificate.

As in the case of the privacy certificate, the transfer agreement should designate the individual or official of the recipient organization who will have primary responsibility for maintenance of transferred data.

A transfer agreement must be entered into prior to transfer of data in identifiable form for subsequent research/statistical uses.

A transfer agreement is not required where data is transferred to a sub-contractor, provided that provisions assuring security and nonrevelation of data, (consistent with requirements of the regulations), are included in the sub-contract agreement.

It is recommended that a copy of the transfer agreements be retained by both the transferor and the recipient of data. Copies of the transfer agreement are not required to be submitted to or approved by LEAA.

Where information is to be retransferred by a recipient of data, the transfer agreement between primary and secondary recipient of data should be reviewed by the original researcher prior to approval of the secondary transfer.

Interface with LEAA Regulations on Criminal History Information Systems (28 CFR Part 20)

The LEAA regulations covering Privacy and Security of Criminal History Information (28 CFR Part 20) provide that agencies covered by the regulations may, but are not required, to disclose identifiable criminal history information for a research or statistical purpose. Such disclosure is permitted regardless of whether or not the proposed research or statistical activity is LEAA supported.

Where criminal history information is released for such purposes, an agreement ensuring confidentiality of the data must be entered into between the criminal justice agency and the recipient of the data. Where data recipients have submitted a privacy certificate in connection with the project for which the criminal history information is to be used, the certificate would be sufficient to fulfill this requirement. If no privacy certificate has been submitted (e.g., if the research is not LEAA-supported), the agreement should contain assurances similar to those required for the privacy certificate.

Agencies subject to 28 CFR Part 20 should note that release of data for a research/statistical purpose does not subject the agency to provisions of the confidentiality regulations (28 CFR Part 22). This is the case since the confidentiality provisions apply only to data which are obtained for research/statistical purposes and not to be basic records from which such data is extracted.

SUGGESTED FORMAT--SAMPLE ONLY

PRIVACY CERTIFICATION

Title of Project

Name of Grantee

The Privacy Certification should contain the following information:

I. A description of the Research/Statistical component of project (or if this information is contained in the grant proposal, a notation of where in the grant proposal the information is located). If questionnaires are to be utilized, attach copy.

II. A justification for collection and/or maintenance of data in identifiable form and description of procedures to be followed to preserve anonymity of private persons as required by Sec. 22.23(b)(7).

III. A description of physical and/or administrative procedures to be followed to insure the confidentiality of data (including procedures for notification of staff and sample staff notification agreement as required by Sec. 22.23(b)(2)).

IV. A description of the procedures to be used for notification of subjects as required by Sec. 22.23(b)(4), or if such notification is to be waived, pursuant to Sec. 22.27(c) a justification therefore.

Where identifiable information is to be used for non-research or statistical purposes, a sample or description of the Consent Statement to be used, shall be attached.

V. A sample of the Transfer Agreement to be used for transfer of data in identifiable form. Indicate the name and title of the individual with the authority to transfer data. Also describe any institutional limitations or restrictions applicable to such transfers.

VI. A description of procedures to be followed for final disposition of data, and where a name index is to be maintained, a description of procedures to secure the index as required by Sec. 22.25(b). Indicate the name and title of the individual authorized to determine the final disposition of data.

The Certification should also contain an assurance such as the following:

Grantee certifies that:

(1) the information contained above is correct and that the procedures noted above will be carried out;

(2) the project will be conducted, consistent with all requirements of Sec. 524(a) of the Omnibus Crime Control Act of 1968, as amended, and Regulations promulgated thereunder contained in 28 CFR Part 22;

(3) LEAA will be notified of any material changes in any of the information supplied above.

Signature of person authorized to sign for grantee

Signature and title of project director or other official primarily responsible for use and maintenance of confidential data (if same as above, indicate)

Date

SUGGESTED FORMAT--SAMPLE ONLY

Information Transfer Agreement

Title of Project for which information was originally compiled, obtained, or used

Name of Individual or Organization to which information is being transferred

LEAA Grant or Contract Number

- Title of Project for which data will be used

The transfer agreement should contain the following information:

I. A description of the Research/Statistical component of the project and a statement of how the project plan will be designed to preserve the anonymity of private persons to whom the information relates.

II. An assurance that the recipient of data is familiar with the Department of Justice regulations, (28 CFR Part 22), and agrees to comply with them.

III. An assurance that information identifiable to a private person that is transferred pursuant to this agreement will be used for research and statistical purposes only and will not be revealed except as allowed under §22.24(b), (c) of the regulations--project findings and reports prepared for dissemination will also not contain such information.

IV. A description of the administrative and physical precautions that will be taken to assure security of information obtained.

V. An assurance that the final disposition of the information transferred has been determined by the parties to this agreement and is in accord with §22.24(h). This should include a description of the procedures.

The recipient agrees that any violation of this agreement will constitute a violation of the Department of Justice regulations, and be punishable as such.

Signature of person authorized to transfer this data

Signature of person receiving data and assuming responsibility for its confidentiality and security

SECTION 5D

EXAMPLES OF FORMS USED TO OBTAIN NAMES OF JUVENILE OFFENDERS FOR THE PURPOSE OF INTERVIEWING THEM

Overview

The forms include a transfer agreement, privacy certificate, employee and subcontractor agreement, informed consent contact letter, and an informed consent agreement. The forms were used to obtain the names of juvenile offenders who had participated in a restitution program. The names were needed so the youths could be contacted and interviewed. A letter (not shown) describing the purpose of the study was sent to the agency from which the data were requested.

CONTINUED

6 OF 7

(INFORMATION TRANSFER)

AGREEMENT BETWEEN

INSTITUTE OF POLICY ANALYSIS

and

SEATTLE COMMUNITY ACCOUNTABILITY PROGRAM

AGREEMENT made this _____ day of _____, 1978, between the Institute of Policy Analysis (hereinafter referred to as IPA) and The Seattle Community Accountability Program (hereinafter referred to as CAP.

WHEREAS, CAP maintains certain files and records in conjunction with its statutory duties and obligations.

WHEREAS, IPA is conducting a national evaluation of juvenile restitution programs;

WHEREAS, in order for IPA to perform its evaluation it is necessary that IPA conduct interviews with youth under the supervision of CAP;

WHEREAS, IPA will use said interview information only for research, evaluative, and statistical purposes in the conduct of its evaluation;

WHEREAS, IPA will not maintain any of the interview information in the juvenile files or records in identifiable form;

WHEREAS, IPA represents, that it is in receipt of, and is familiar with the provisions of 28CFR Part 22, including provisions for sanctions set out at Section 22.29;

NOW THEREFORE, IT IS AGREED AS FOLLOWS:

(1) CAP will grant IPA access to selected youth who consent to be interviewed in regard to their offense histories, perceptions

of offense sanctions, and experiences with restitution requirements;

(2) CAP will provide names of youths who have indicated they might be willing to participate in an interview to _____ so that _____ can obtain informed consent from the youths and administer the interview questionnaire.

(3) IPA will:

(a) Use the said information only for research, evaluative, and statistical purposes in conducting its national evaluation of juvenile restitution programs and for no other purpose.

(b) Limit access to said information to employees or subcontractors of IPA whose responsibilities cannot be accomplished without access to the data and who have agreed in writing to comply with the provisions of this agreement, the privacy certificate, and the provisions of 28 CFR part 22;

(c) Store all information received pursuant to this agreement in secure locked containers;

(d) Identify interview respondents with a numeric or other appropriate code;

(e) Immediately notify CAP in writing of any proposed material changes in the purposes or objectives of its research, or in the manner in which that information will be used.

(4) IPA will not:

(a) Disclose any of said information in a form which is identifiable to an individual in any project findings or reports, or in any manner inconsistent with the provisions of 28CFR part 22.

(b) Copy any of said information, except as clearly

necessary for use by employees or contractors to accomplish the purposes of the research;

(5) In the event that IPA deems it necessary for purposes of the research to disclose said information to any subcontractor other than _____, IPA shall obtain prior written consent from CAP and shall also secure the written agreement of the subcontractor to comply with the terms of this agreement as if it were named here;

(6) IPA further agrees that:

(a) CAP shall have the right, at any time, to monitor, audit, and review the activities and policies of IPA or its subcontractors in implementing this agreement in order to insure compliance therewith;

(7) In the event IPA fails to comply with any terms of this agreement CAP shall have the right to take such action as it deems appropriate including termination of this agreement. If CAP terminates this agreement, IPA and any subcontractor shall forthwith return all said information and all copies made thereof to CAP or make such alternative dispositions thereof as directed by CAP. The exercise of remedies pursuant to this paragraph shall be in addition to, and not limit any other sanctions provided by law or other legal remedy available to parties injured by unauthorized disclosures of juvenile record information.

(8) IPA will hold CAP harmless from any damages or other liability which might be assessed against CAP as a result of disclosure by IPA or any of its subcontractors, of any information received pursuant to this agreement.

IN WITNESS WHEREOF, the parties have signed their names hereto this _____ day of _____, 1978.

SEATTLE COMMUNITY ACCOUNTABILITY PROGRAM

By_____

Title_____

INSTITUTE OF POLICY ANALYSIS

By_____

Director of Administration

PRIVACY CERTIFICATE

INSTITUTE OF POLICY ANALYSIS
National Evaluation of Juvenile Restitution Program

I.

The Institute of Policy Analysis (IPA) has received a grant from the Office of Juvenile Justice and Delinquency Prevention to conduct a national evaluation of juvenile restitution programs. The purpose of this evaluation is to test the effectiveness of experimental restitution programs in a juvenile court setting in terms of reducing recidivism of juvenile offenders and increasing victim satisfaction with the juvenile justice system.

II.

In order to conduct the evaluation, it will be necessary to obtain information from juveniles in the form of interviews conducted by IPA research staff or subcontractors. Prior to the interviews, written consent to participate will be obtained from all juveniles and their parents or legal guardians. The written consent will include a description of the purpose of the interview, procedures to protect the confidentiality of the information obtained, and the stipulation that subjects may withdraw from the interview at any time without penalty. The data we collect will only be used for research, evaluative, and statistical purposes. The project findings, and reports prepared for dissemination will contain no information which can reasonably be expected to be identified to a private person.

III.

The following physical and administrative procedures will be followed to insure that the confidentiality of data is maintained:

1. All data will be stored in locked filing cabinets.

2. Access to the data will be limited to employees or subcontractors of IPA who have a need to use the data and who have agreed in writing to comply with IPA's Privacy Certificate.

IV.

The data which have been collected may only be transferred pursuant to a written transfe r agreement and only the project director and the director of administration have authority to authorize the transfer of data.

IPA certifies that:

1. The information contained above is correct and procedures noted above will be carried out;

2. The research project will be conducted consistent with all requirements of Sec. 524(a) of the Crime Control Act of 1973 as amended and regulations promulgated thereunder contained in 28CFR Part 22;

3. OJJDP will be notified of any material changes in any of the information supplied above.

INSTITUTE OF POLICY ANALYSIS

By_____
Director of Administration

By_____
Project Director

Date_____

EMPLOYEE AND SUBCONTRACTOR AGREEMENT

I, _____,

Subcontractor/Employee of the INSTITUTE OF POLICY ANALYSIS, acknowledge familiarity with the IPA data transfer agreement, the Privacy Certificate concerning confidential data, and the provisions of 28CRF Part 22 and agree to comply with the terms and conditions thereof in my use and protection of the juvenile justice information obtained pursuant to the written transfer agreement.

_____

DATE_____ _____

(INFORMED CONSENT CONTACT LETTER)

Dear

We are currently in the process of evaluating a demonstration juvenile program throughout the United States. Part of the evaluation includes a personal interview with certain young persons in King County. a researcher for this project, will be conducting the interview.

She will be calling within the next few days to explain the research and to set up an appointment with (name of child) . The interview is voluntary. (Name of child) does not have to consent to be interviewed and can withdraw at any time even after it starts. Youths who participate in the research will be paid $5.00. The information will be completely confidential and will be reported as group information with no individual names attached. Individual questionnaires will be destroyed.

We hope (name of child) will assist us in our evaluation.

Sincerely,

INFORMED CONSENT FORM FROM

THE INSTITUTE OF POLICY ANALYSIS

for the

JUVENILE RESTITUTION EVALUATION

I, _____ agree to be interviewed by the Institute of Policy Analysis (IPA) for the purpose of discussing my attitudes and my experiences, including those with the Seattle Community Accountability Program. I understand that even if I agree to the interview, I may withdraw my permission at any time without penalty. I understand that I will be compensated in the amount of $5.00 if I promptly complete the interview. I understand that IPA is interviewing me to help learn whether restitution and/or community service are worthwhile programs for juveniles. I understand that all information obtained will only be used for research and statistical purposes, that all information collected will be kept confidential, and that my name will not be used or revealed in any way that can identify me with the responses I have given to the interviewer.

Date_____ _____
(Juvenile)

_____
(Witness)

Date_____ _____
(Parent or Guardian)

_____
(Witness)

SECTION 5E

EXAMPLE OF FORM

TO BE USED WITH HUMAN SUBJECTS REVIEW COMMITTEE

Overview

This form contains the questions for which answers are needed in order to comply with the Department of Health, Education and Welfare requirements for the protection of human subjects. It is filled out by the evaluator (or other researcher) and submitted to a Human Subjects Review Committee along with a copy of the grant application.

HUMAN SUBJECTS REVIEW FORM

1. **Briefly** describe the nature of your Proposal, including the activities **involving** human subjects. Your written description is to be one page or **less.** If a one page or less description is already included in your **Proposal,** you need only identify the page in your Proposal where it may **be found.**

2) Describe the characteristics of the human subject group of groups involved in your Proposal (See Guide, p.1, paragraph B).

A) Sex, race or ethnic group, age range, etc.

B) Affiliation of subjects, e.g., institutions, hospitals, general public, etc.

C) Subjects' general state of health.

3) If human subjects are either children, mentally incompetent or legally restricted groups, give explanation as to: A) The necessity for using these particular groups; and B) Why adult "normal" groups cannot be used (specifically).

4) A) Describe the known or foreseeable risks of harm (physical, psychological, economical, sociological, legal, privacy, confidentiality or sensitive information, or other) to which the human subjects will be exposed, both immediate and long range. There are possible risks involved with interviews, questionnaires, tape recordings, photographs, field work, work with children, evaluations, deception, final research publications, etc. (See Guide at pages 2-3, paragraphs C and D).

a) Immediate risks.

b) Long range.

c) Rationale for the necessity of such risks.

d) Alternatives that were or will be considered.

e) Why alternatives may not be feasible.

B) Describe the safeguards to be used to eliminate or minimize each of the possible risks of harm stated in 4)A) above, e.g., anonymity, security, code lists, use of physicians, psychologists, etc., and what precautionary measures will be taken to insure the protection of human subjects, e.g.:

a) Type of consent to be obtained (written or oral).

b) How and where will permission be recorded.

c) If subjects are minors or mentally incompetent, describe how and by whom permission will be granted.

If employing the use of questionnaires or other survey instruments, include a copy of each sample.

5) What precautions will be taken to safeguard identifiable records of individuals? These questions also apply if you are using secondary sources of data:

A) Consider both the long range use and immediate use of data (by you and others).

B) Describe specific procedures to be used to provide confidentiality of data.

6) A) Informed consent is required for subjects at risk. A sample of an Informed Consent Form should be prepared and submitted for the Committee's review. Note that the Consent Form must include the following:

a) A fair explanation of the procedures to be followed, including an identification of those which are experimental;

b) A description of the attendant discomforts and risks;

c) A description of the benefits to be expected;

d) In client service situations a disclosure of appropriate alternative procedures that would be advantageous for the subject;

e) An offer to answer any inquiries concerning the procedures; and

f) An instruction that the subject is free to withdraw his consent and to discontinue participation at any time.

(See Guide at p.7, paragraph (c).)

B) If the subjects are minors a parent consent form is required (See Guide p.7, paragraph C).

C) In special and/or unusual circumstances, consent of the subjects can be obtained orally. If obtained orally, explain why it is not obtained in writing, and submit a copy of the information to be given orally. (See Guide at pp 7-8 paragraph C.)

7) Describe the potential benefits to the subjects (See Guide at pp.6-7, paragraph (b)).

8) Describe the potential benefits to humanity / e.g., the importance of the knowledge to be gained, and the usefulness of the information gained to the community at large. (See Guide at pp.6-7, paragraph (b).)

9) "Non-Beneficial Research" is defined as research involving physiological and psychological investigations of a person, his body or surroundings, which is devoid of therapeutic purpose to that person.. If you plan to conduct this type of research and feel that there are no other methods available for obtaining the information needed, please describe:

A) What other methods were or will be explored.

B) The extent of the risks (Describe in detail any physical, psychological, social, legal and other risks you can foresee both immediate and long range).

C) The importance of the knowledge to be gained.

D) Why you feel that the value of information to be gained outweighs the risks.

PART II

DEVELOPMENT OF CONTINUING RESOURCES FOR EVALUATORS

SECTION 6

OVERVIEW OF PART II

OVERVIEW OF PART II

Introduction

This portion of the handbook is divided into four sections: (a) reference services and bibliographic materials, (b) computer resources, (c) data sources, and (d) organizational and administrative information on each of the eight regional planning units in which an evaluator is employed.

Users of these materials should be reminded that one purpose of the handbook is to provide both current and future evaluators with a guide to the available resources existing in their RPUs. Each evaluator, therefore, should undertake the responsibility of updating these materials periodically to ensure the currency of the information. This is especially important for those evaluators intending to vacate their positions in the near future, as the new information which they have acquired should not be lost, but rather passed on to their successors.

Reference Services and Bibliographic Materials

This section contains the names and addresses of four national reference services in the areas of criminal justice and evaluation research. Also included are the instructions for using the services and samples of the types of materials which may be obtained from them.

The bibliography contained in this section is annotated and includes major works in the fields of evaluation research and methodology. Users can and should expand this section of the Handbook by adding bibliographies already in their possession and by enlarging the existing list with

citations of works they have found particularly useful.

Computer Resources*

The section on computer resources contains all the information needed to gain access to and use the computing facilities at the University of Washington, Eastern Washington State College, and Western Washington University. While the information contained in this section is up-to-date as of August 1978, the reader should be aware that changes undoubtedly will occur as equipment at the computing centers is replaced, new programs become available, and so forth. Frequent users of the facilities will have no trouble keeping pace with new developments, however, as computing centers invariably issue or post memoranda announcing changes in procedures or the availability of new programs. The inclusion of these memoranda in this Handbook would be an excellent way of keeping current.

Also included in this section is a set of instructions for using an interrupted time series program which may be accessed from a remote terminal at the University of Washington, and a sample of the statistical packages available at Eastern Washington State College. Evaluators could expand upon this part of the Handbook by adding similar materials from whichever computing facilities they use.

Data Sources*

In this section are the major sources of data as reported by each of the regional planning units in which an evaluator is employed. With but a few exceptions, local criminal justice data are not routinely

*This is one of the two sections omitted from the copies distributed by LEAA because it pertains only to the State of Washington.

maintained in the storage banks of computers; rather, the evaluator usually must seek out the persons responsible for maintaining records, gain access to those records, and code the data manually. The names and/or titles of persons responsible for maintaining records in each of the RPUs are included in this section. This section should be updated by the local evaluator each time s/he collects data for a research project.

Organizational and Administrative Information

This section contains organizational charts and brief narratives explaining the administrative structure of each RPU and how the office fits into the local governmental system. Also included are the names of the members of the local Law and Justice Advisory Committees, which usually are the bodies responsible for making funding decisions. As the membership of these committees changes with expirations of terms, resignations and so forth, the list should be updated every several months.

SECTION 7

SOURCES OF INFORMATION

FOR CRIMINAL JUSTICE EVALUATORS

SOURCES OF INFORMATION FOR CRIMINAL JUSTICE EVALUATORS

Within recent years computerized reference services have become valuable technical information sources for criminal justice researchers. Four computerized bibliographic reference services are listed below which should be of particular value to Law and Justice evaluators.

National Criminal Justice Reference Service

The National Criminal Justice Reference Service (NCJRS) is perhaps the most widely used bibliographic information service. If one wishes to obtain an annotated listing of recent research on citizen involvement in crime prevention programs (for example), one simply write NCJRS at the address below and requests a computerized list of references on that particular topic. More than one topic may be requested at a time and the NCJRS is very good at including similar (or alternative) topics which might be related to the topic one has requested. In the request letter one must specify the LEAA grant or project under which one is working. Attachment 1 contains an example of materials produced by NCJRS. The user is not billed for this service.

National Criminal Justice Reference Service
Post Office Box 24036
Washington, DC 20024
(202) 862-2900

National Clearinghouse for Mental Health Information

The National Clearinghouse for Mental Health Information is located within the National Institute for Mental Health. In general, one will receive somewhat different materials from the Clearinghouse than from

NCJRS for a similar requested topic because different agencies often provide materials to only one of these two agencies. For example, information on drug related crime would probably be more plentiful at the Clearinghouse than at NCJRS, while information on the role of police in school desegregation would be more accessible at NCJRS. A request for materials from the National Clearinghouse should be made to the address below. Again, more than one topic can be requested at a time. Be sure to include the LEAA grant number or project name under which the research is being conducted. The user is not billed for this service. Attachment 2 shows an example of materials from the Clearinghouse.

Technical Information Center
National Clearinghouse for Mental Health Information
National Institute for Mental Health
Room 15-C-26
5600 Fishers Lane
Rockville, Maryland

Databank of Program Evaluations

The Databank of Program Evaluations (DOPE) is located at the UCLA School of Public Health. DOPE is available in both an on-line interactive mode and an off-line mode. Users are charged a use fee, usually from $20 to $30, depending on the number of references available under the topic requested. Moreover, the scope of topics is more limited than it is with NCJRS and the Clearinghouse and the citations are often from journals.

More information on DOPE is available by writing to the address below. Attachment 3 contains an example of materials from this service.

UCLA Databank of Program Evaluations
UCLA School of Public Health
University of California, Los Angeles
Los Angeles, CA 90024

Center for Law and Justice, University of Washington

The Center for Law and Justice has computerized bibliographic citations available on topics relating to the prevention of juvenile crime. The Center's reference service is funded through the Office of Juvenile Justice and Delinquency Prevention (OJJDP). At present the citations are still being compiled and when they are completed they will be included in the National Criminal Justice Reference Service's inventory and accessible through NCJRS. In the interim, requests to the Center are being honored on various topics concerning juvenile crime prevention. The complete address and contact person are listed below.

Ms. Janette Schueller
Center for the Assessment of Delinquency Behavior
and Its Prevention
Center for Law and Justice, JD-45
University of Washington
Seattle, WA 98195
(206) 543-1485

Attachment 1. An Example of NCJRS Materials

D043 01/16/78 16:23 PAGE 173

SET/40 DOCUMENTS 1:105

DOCUMENT 28

ACCESSION NUMBER:.... 09900.00.009445
TITLE: POLICE PROGRAMS FOR PREVENTING CRIME AND DELINQUENCY
PUBLICATION DATE: 72 PAGES: 509
AUTHOR(S): PURSUIT, D. G. GERLETTI, J. D.
SALES AGENCY: CHARLES C THOMAS
301-327 EAST LAWRENCE AVENUE
SPRINGFIELD IL 62717

ANNOTATION:
SEVENTY JOURNAL ARTICLES, PROGRAM DESCRIPTIONS, AND OTHER MATERIAL ON THE POLICE AS CRIME CONTROLLING AGENTS.

ABSTRACT:
A VARIETY OF PERCEPTIONS ON THE ROLE OF LAW ENFORCEMENT IN THE PREVENTION OF CRIME AND DELINQUENCY ARE PROVIDED ALONG WITH A HISTORICAL PERSPECTIVE. A NUMBER OF REPRESENTATIVE COMMUNITY RELATIONS PROGRAMS ARE ILLUSTRATED, ALONG WITH PROJECTS AIMED AT PREVENTING SPECIFIC OFFENSE SUCH AS AUTO THEFT, BURGLARY, CHILD MOLESTATION, ROBBERY, JUVENILE DRUG USE, AND ASSAULTS ON POLICE IN FAMILY CRISES. VARIOUS SCHOOL RELATED PROGRAMS DESIGNED TO REACH YOUTH AT AN EARLY AGE ARE COVERED, AS WELL AS RECREATIONAL AND LEISURE TIME PROGRAMS ORGANIZED BY POLICE TO CURE DELINQUENCY. ALSO INCLUDED IS A REVIEW OF THE EFFECTIVE USE OF COMPUTERS AND HELICOPTERS BY LAW ENFORCEMENT AGENCIES. THE VOLUME CONCLUDES WITH A DESCRIPTION OF THE FUNDING RESOURCES OF THE LEAA AND THE YOUTH DEVELOPMENT AND DELINQUENCY PREVENTION ADMINISTRATION ALONG WITH GUIDELINES FOR PREPARING GRANT PROPOSALS AND DEVELOPING EVALUATION PROCEDURES. BECAUSE OF ITS WIDE SCOPE, THIS BOOK SHOULD BE OF INTEREST TO MANY INDIVIDUALS IN THE LAW ENFORCEMENT AND CRIMINAL JUSTICE COMMUNITY. (SNI ABSTRACT)

DOCUMENT 29

ACCESSION NUMBER:.... 09900.00.009596
TITLE: POLICE-COMMUNITY RELATIONS - A PRACTICAL GUIDE FOR TEXAS POLICE OFFICERS
PUBLICATION DATE: 70 PAGES: 215
AUTHOR(S): BERTOTHY, B
CORPORATE AUTHOR: TEXAS COMMISSION ON LAW ENFORCEMENT OFFICER STANDARDS AND EDUCATION
503-E SAM HOUSTON BUILDING
AUSTIN TX 78701

ANNOTATION:
DESCRIPTION OF COMMUNITY RELATIONS PROGRAMS, INCLUDING THEIR PURPOSES, THEIR INSTITUTIONS, AND DESIGN, AND SPECIFIC ILLUSTRATIONS OF PROGRAMS

Attachment 2. An Example of National Clearinghouse Materials

W-4769, BURCART, CRIME;

76-2731 L5
AUTHORS: Hamill, Pete.
ADDRESS: New York Post, New York, NY
TITLE: The porno war.
SOURCE: In: Kaplan, L., An economic analysis of crime: selected readings.
SOURCEID: Springfield, IL, Charles C Thomas, 1976. 410 p. (p. 323-325).

The success of the current crackdown on pornography and prostitution is discussed. Prostitution and pornography are distinguished from the "true crimes" of violence, poverty, drug addiction, and street crimes. It is contended that prostitution should be legalized instead of attempting to enforce the moral judgments of the law. It is suggested that prostitution and pornography will continue to exist despite legal attempts to curb or to end it and, therefore, it should be made available to those who care to use it.

76-4159 L5
AUTHORS: Mann, Fredrica; Friedman, C. Jack; Friedman, Alfred S.
ADDRESS: Philadelphia Psychiatric Center, Philadelphia, PA
TITLE: Characteristics of self-reported violent offenders versus court identified violent offenders.
SOURCE: International Journal of Criminology and Penology (London).
SOURCEID: 4(1):69-87, 1976.

A study exploring differences between youths apprehended by police for commission of violent acts and youths who reported the commission of violent acts but had no official record for such behavior is reported. Comparisons were based on extensive tests and questionnaires yielding psychological, sociological, demographic, family background and interaction, and legal data. The study combined multiple relevant factors within a single comparative procedure of analysis. Hierarchical structure of variables highly associated with each offender group was also determined. Subjects were 61% Black, 39% White, boys between 15 and 18 years old, from poor and disadvantaged families, and with police records in 66% of the cases. Criteria found related to self-report violence included parental defiance, negative family role behavior, street gang membership, drug abuse, youngest age delinquency, and alcohol abuse. Court record violence criteria correlates were negative family role behavior, street gang membership, disruptive family role behavior, mother's managerial family role behavior, unrealistic level of aspiration, and scores on Gorham multiple choice proverbs test. It is concluded that for both groups of offenders the most powerful predictors of their violent behavior were family related measures. 12 references.

Attachment 3. An Example of Databank of Program Evaluations Materials

D O P E Offline Listing Page 68

Document ID#1028 DOPE Filing #3062
Author(s): MARVIT, ROBERT C. LIND, JUDY MCLAUGHLIN, DENNIS
Title: USE OF VIDEOTAPE TO INDUCE ATTITUDE CHANGE IN DELINQUENT ADOLESCENTS
Journal: AMERICAN JOURNAL OF PSYCHIATRY. VOLUME 131, ISSUE 9 (SEPTEMBER 1974) PAGES 996-999.

Condition: ANTISOCIAL BEHAVIOR IN ADOLESCENTS

Sample Description --
Age: ADOLESCENTS
Sex: 61% FEMALE
Race: UNSPECIFIED
Income: UNSPECIFIED
Sample Size: 44

Type of Treatment: GRP TRT W/ & WITHOUT VIDEOTAPING. EXP GRPS (N=23) VIDEOTAPED DURING 4 PSYCHOTHERAPY SESSIONS. GRPS VIEW & DISCUSS TAPES, INCLUDING FEELINGS ABOUT SEEING THEMSELVES & HOW SELF-IMAGE COMPARED TO VIDEOTAPE

Site: RESIDENTIAL

Study Design: CONTROLS W/ RANDOM ASSIGNMENT, BEFORE-AFTER MEASURES, STATISTICAL TESTS

What is Measured 1: ALIENATION/ATTITUDE
Measures Used: UNSPEC (SELF-REPORTING ON QUESTIONNAIRE ITEMS BEFORE & AFTER TRT)
Outcome: SIG INCR FOR VIDEOTAPE GRP (N=23) IN FEELINGS OF GETTING ALONG W/ OTHERS. SIG DIFF FROM CONTROL (N=21)
Data Collection: SELF-REPORTING, ADMINISTRATION OF QUESTIONNAIRES

What is Measured 2: SELF CONCEPT
Measures Used: UNSPEC (SELF-REPORTING ON QUESTIONNAIRE ITEMS BEFORE & AFTER TRT)
Outcome: SIG INCR FOR VIDEOTAPE GRP (N=23) IN FEELINGS OF DIMINISHED SELF-CONCEPT AS IT RELATES TO OTHERS. SIG DIFF FROM CONTROL (N=21)
Data Collection: SELF-REPORTING, ADMINISTRATION OF QUESTIONNAIRES

What is Measured 3: SOCIAL BEHAVIOR
Measures Used: UNSPEC (MODIFICATION OF DRESS AFTER VIDEOTAPINGS)
Outcome: MODIFICATION OF BEHAVIOR TO IMPROVE APPEARANCE WAS REFLECTED BY CHANGE OF DRESS IN VIDEOTAPE GRP
Data Collection: SELF-REPORTING

Conclusions: NO ONE REFUSED TO PARTICIPATE IN THE STUDY DESPITE UNRESPONSIVENESS OF MANY DELINQUENT YOUTHS. THE VIDEOTAPING EXPERIENCE WAS POSITIVE IN THAT YOUTHS SAW THEMSELVES AND THEIR BEHAVIOR IN A UNIQUE AND FASCINATING WAY. AFTER THE FIRST REPLAY, ALL SEEMED DISAPPOINTED IN HOW THEY LOOKED AND SOUNDED. AFTER DISCUSSION OF FEELINGS AND REACTIONS, THE QUESTIONNAIRE DEMONSTRATED THEY HAD MORE COMFORTABLE FEELINGS ABOUT THEIR APPEARANCE. THE SITUATION WAS ANXIETY PROVOKING FOR SEVERAL.

Program Address: DR. MARVIT, MENTAL HEALTH RESEARCH TEAM, P.O. BOX 3378, HONOLULU, HAWAII 96801

A BIBLIOGRAPHY
OF THE BASIC LIBRARY
FOR AN EVALUATOR

Introduction

The bibliography presented below is intended to serve as a "first reference" for evaluators who may seek citations when confronting a particular problem. No attempt has been made to present an exhaustive bibliography for any of the topic headings listed. Instead, we have reflected upon the books and articles that we have found particularly valuable in undertaking evaluative research and these are listed below.

In some instances, personal preference for style of presentation may lead one to adopt one source over another (this is particularly true for choice of a statistics text). In others, someone may prefer a book that is not listed to one that we have presented below. We have sought, with the exception of a statistic text, to provide what we think to be the best treatment of a particular issue, recognizing that others, perhaps equally good or readable, may exist.

Evaluation Research: General

Marcia Guttentag and Elmer L. Struening, eds., HANDBOOK OF EVALUATION RESEARCH volumes 1 and 2 (Sage Publications, 1975). These two volumes contain numerous articles of interest to those involved in evaluative research. The content of the articles range from theories of evaluation, through the methodology of evaluation. Among other topics covered are those of cost-benefit and cost-effectiveness evaluation.

Edward A. Suchman, EVALUATIVE RESEARCH: PRINCIPLES AND PRACTICE IN PUBLIC SERVICE & SOCIAL ACTION PROGRAMS (Russell Sage Foundation, 1967). Something of a standard introduction to the theory and practive of evaluative research. Includes discussion of both methods of evaluation and the use of evaluation in the bureaucratic setting.

Daniel Glaser, ROUTINIZING EVALUATION: GETTING FEEDBACK ON EFFECTIVENESS OF CRIME AND DELINQUENCY PROGRAMS (Department of Health, Education and Welfare, 1973). A reasoned discussion of the application of different types of evaluation designs to different kinds of evaluation problems.

Lee R. McPheters and William B. Stronge, eds., THE ECONOMICS OF CRIME AND LAW ENFORCEMENT (Charles C. Thomas, 1976). An excellent collection of articles written by economists with a focus upon the criminal justice system. For those whose academic background is largely in psychology or sociology, the articles contained in this collection will be found to be argumentative.

Scarvia B. Anderson and Samuel Ball, THE PROFESSION AND PRACTICE OF PROGRAM EVALUATION (Jossey-Bass Publishers, 1978). This is an excellent book for the consumers of evaluations as well as for the producers of evaluations. Of particular interest are discussions concerning the most appropriate methods for the different purposes of evaluation and a chapter on the problems of training and assessing the skills of evaluators.

Scarvia B. Anderson and Claire D. Coles (eds.), EXPLORING PURPOSES AND DIMENSIONS (Jossey-Bass Publishers, 1978). This is the first volume in a series entitled "new directions for program evaluation." The book is a reader containing articles by a variety of authors on different aspects of program evaluation.

Time-Series Analysis

Charles W. Ostrom, Jr., TIME-SERIES ANALYSIS: REGRESSION TECHNIQUES (Sage Publications, 1978). A good and readable review of the application of regression analysis to time-series analysis.

George E.P. Box and Gwilym M. Jenkins, TIME SERIES ANALYSIS: FORCASTING AND CONTROL revised edition (Holden-Day, 1976). A highly technical but authoritative treatment of time-series analysis.

Gene V. Glass, Victor L. Willson, and John M. Gottman, DESIGN AND ANALYSIS OF TIME-SERIES EXPERIMENTS (Colorado Associated University Press, 1975). Apparently the first attempt to apply the Box and Jenkins time-series models to the investigation of interrupted time-series data. This is also a highly technical presentation.

Warren Gilchrist, STATISTICAL FORCASTING (John Wiley & Sons, 1976). A much more readable, although still technical, presentation of the Box and Jenkins time-series models.

Charles R. Nelson, APPLIED TIME SERIES ANALYSIS (Holden-Day, Inc., 1973). Like Gilchrist, this is a more readable, although technical, presentation of the Box and Jenkins time-series models.

Design

Donald T. Campbell and Julian C. Stanley, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR RESEARCH (Rand McNally & Company, 1966). Probably the most frequently cited authority on the utility of various research designs.

Thomas D. Cook and Donald T. Campbell, "The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings," in HANDBOOK OF INDUSTRIAL AND ORGANIZATIONAL RESEARCH (Rand McNally, 1975). An extension of the work on research designs first reported by Campbell and Stanley. An excellent supplement to that earlier work.

Leslie Kish, SURVEY SAMPLING (John Wiley & Sons, Inc., 1965). Generally considered the most authoritative treatment of sample theory and practice currently available.

Eugene J. Webb, Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest, UNOBTRUSIVE MEASURES: NONREACTIVE RESEARCH IN THE SOCIAL SCIENCES (Rand McNally & Company, 1973). Along with being a very engaging discussion of social science research methods, it is a book which should provoke the reader to consider inventive ways of obtaining data without biasing the response of research subjects.

Donald T. Campbell and Robert F. Boruch, "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects," in C.A. Bennett and A. Lumsdaine (eds.), CENTRAL ISSUES IN SOCIAL PROGRAM EVALUATION (Academic Press, 1975). A cogent but rather technical discussion of the problems in overcoming selection bias in non-randomly selected comparison and treatment groups. The authors argue persuasively for random assignment.

Measurement

Robyn M. Dawes, FUNDAMENTALS OF ATTITUDE MEASUREMENT (John Wiley & Sons, 1972). This is an excellent introduction to attitude measurement. The author outlines the theory of attitude measurement and a variety of approaches that may be adopted.

J.P. Guilford, PSYCHOMETRIC METHODS (McGraw-Hill Book Company, 1954). This is something of a standard in the area of attitude measurement. It is more inclusive than the Dawes book, but is also considerably more difficult.

William A. Mehrens and Robert L. Ebel, eds., PRINCIPLES OF EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT: A BOOK OF SELECTED READINGS (Rand McNally & Company, 1967). This is an excellent compendia of essential articles in the area of attitude measurement.

Statistics

Linton C. Freeman, ELEMENTARY APPLIED STATISTICS: FOR STUDENTS IN BEHAVIORAL SCIENCE (John Wiley & Sons, Inc., 1968). A good introduction to notions of association and significance. The presentation of relevant statistics is particularly appropriate for those seeking an introduction to the area.

Hubert M. Blalock, SOCIAL STATISTICS (McGraw-Hill Book Company, 1972). This is a standard statistical text for those in the behavioral sciences. Most of the statistical problems one normally encounters in the social sciences (excluding those of time-series analysis) are covered by this text or those of Edwards and Hays, noted below. Little separates the utility of these three texts, in the general case, other than personal preference for style of presentation.

Allen L. Edwards, STATISTICAL METHODS (Holt, Rinehart and Winston, Inc., 1967). See Blalock for discussion of selection of statistics text.

William L. Hays, STATISTICS FOR PSYCHOLOGISTS (Holt, Rinehart and Winston, 1963). See Blalock for discussion on choice of statistics text.

Sidney Siegel, NONPARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES (McGraw-Hill Book Company, Inc., 1956). This still remains as probably the best book available for discussion of the use of non-parametric statistics. Since much of the data used by social scientists is non-parametric, this should be a valued and much used addition to a library.

Eric A. Hanushek and John E. Jackson, STATISTICAL METHODS FOR SOCIAL SCIENTISTS (Academic Press, 1977). This provides an excellent discussion of bivariate and multivariate regression analysis, paying particular attention to the different effects that violation of the assumptions of the model will have on one's findings. Much of the discussion is rather technical.

☆U.S. GOVERNMENT PRINTING OFFICE: 1979-311-379/1683

U.S. DEPARTMENT OF JUSTICE
LAW ENFORCEMENT ASSISTANCE ADMINISTRATION
WASHINGTON, D.C. 20531

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF JUSTICE
JUS-436

SPECIAL FOURTH-CLASS RATE BOOK

END