Patricia F. Waller (1974). North Carolina Symposium on Highway Safety (Vol. 10). Highway Safety Programs: How Do We Know They Work? Chapel Hill, NC: University of North Carolina Highway Safety Research Center.

th
carolina
symposium
on highway
safety

chapel hill, n. c.

spring 1974 ● volume ten ● edited by Patricia F. Waller

Highway Safety Programs: how do we know they work?

Noel F. Kaestner
H. Laurence Ross

Highway Safety Programs:

how do we know they work?

North Carolina Symposium

on highway safety

volume ten

Highway Safety Programs: how do we know they work?

Noel F. Kaestner — Oregon Traffic Safety Commission

H. Laurence Ross — University of Denver

NORTH CAROLINA SYMPOSIUM ON HIGHWAY SAFETY

Raleigh, N.C.

Volume ten **Spring 1974**

The University of North Carolina
Highway Safety Research Center
Chapel Hill, N.C. 27514
B. J. Campbell, Director

A FEW WORDS ABOUT THE SYMPOSIUM TOPIC . . .

Would you back a highway safety program even if it saves only one life?

The answer to this question should be, "No" . . . if there are other programs that can save more lives for the same cost.

Because there are so many highway safety programs competing for support, and because there are limited funds with which to support any of them, there is a responsibility to choose those programs that will save the most lives for the money available. To make the best choices, each program's effectiveness must be measured objectively. Ideally, evaluation should be included in the planning stages. This is not always possible, and many programs, therefore, can only be evaluated long after they have been in operation. In developing new programs or new approaches in existing programs, careful consideration should be given to the evaluation of the program's effectiveness. No highway safety program should ever be considered permanently established. Instead, programs should be modified as evaluations provide new information.

Grateful acknowledgment is made
of the special support provided
by the Highway Users Federation
for Safety and Mobility for this
session of the North Carolina
Symposium on Highway Safety.

TABLE OF CONTENTS

LIST OF FIGURES AND ILLUSTRATIONS

Ross

List of Figures (cont'd)

ABOUT THE CENTER . . .

At the request of the Governor of North Carolina, the 1965 North Carolina State Legisture provided for the establishment of the University of North Carolina Highway Safety Research Center. Dr. B.J. Campbell, then Head of the Accident Research Branch of Cornell Aeronautical Laboratory, was invited to return to his alma mater to direct the new Center. He accepted, and in 1966 the Center officially began operation. Since then the staff has grown to more than fifty, representing skills in experimental psychology, clinical psychology, mathematics, transportation engineering, computer systems, journalism, library science, biostatistics, graphic arts, epidemiology, experimental statistics, general engineering, human factors engineering, and health administration. The University of North Carolina Highway Safety Research Center is the first institution in the South devoted exclusively to research in highway safety.

ABOUT THE SYMPOSIUM . . .

The North Carolina Symposium on Highway Safety is a semi-annual event sponsored by the North Carolina State University School of Engineering, the University of North Carolina School of Public Health, and the University of North Carolina Highway Safety Research Center. This spring's session received special support from the Highway Users Federation for Safety and Mobility. First held in the fall of 1969, the Symposium has three major purposes.

First, it is designed to attract students to acquaint them with the problems and possibilities for research in the field of highway safety.

Second, it is a means of bringing together professional workers in the greater North Carolina area whose interests are related to this field.

And third, the published papers from the Symposium provide on a regular basis major positions and summaries of research in the field of highway safety. It is hoped that these volumes will provide ready resource material for persons interested in this field.

INTRODUCTION

Most highway safety programs have developed over the years independent of sound information on which to base decisions. This statement does not comprise an indictment of highway safety officials, for indeed there has not been available the information that these officials need to make the decisions that they are called upon to provide on a day to day basis. It has been only in recent years, and most significantly since the Federal Highway Safety Act of 1966, that a concerted effort has been made to generate and disseminate data that can be used in developing and implementing highway safety programs.

In highway safety, as in other fields, there has developed a large bureaucracy that is charged with the responsibility for administering and maintaining the highway safety activities that have been generated over the years. At this late stage the researchers and evaluators of programs have entered the scene and begun to interject their enlightening bits of advice. As one motor vehicle administrator put it, "You researchers are always telling us about how our programs do not work, but you never have anything constructive to offer instead." His words contain more truth than we would like to acknowledge. Still, administrators perhaps more than researchers are concerned about how to get the most from the highway safety dollar. This symposium deals with the rather delicate subject of evaluating highway safety programs.

The two participants in this symposium represent the best in evaluation research. Dr. Noel F. Kaestner is well-known and respected for the work he has conducted over the years in the state of Oregon where he has established and maintained an enviable relationship with its Traffic Safety Commission. Dr. H. Laurence Ross is well-known for his publications concerning the effects of the Connecticut crackdown on speeding, and more recently on the effect of the British Road Safety Act of 1967 concerning drinking drivers.

Dr. Kaestner's comments on evaluation focus on driver improvement programs. He outlines what he calls a full-scale driver improvement program in which the first step occurs when the driver licensing agency sends an advisory letter

warning the recipient to improve his ways. Should the letter not prove effective, there follows an invitation to appear for an interview with an official of the agency. Should the driver still fail to show improvement, the last step is suspension or revocation of his license. While most evaluative studies have focused on the interview portion of the program, there are some reports of evaluations of each of these phases.

However, the road to effective evaluation of such programs is fraught with many potholes. The first and perhaps the major problem concerns whether anyone really wants to know if programs work. Some do not want to know because they are convinced the programs are effective. Others have strongly vested interests in the programs and do not want to consider the possibility of change. Still others do not want to evaluate programs because they are already convinced the programs are not worthwhile. They can point out that while treated groups of poor drivers invariably show improvement, such improvement can be accounted for by the phenomenon called regression to the mean. Any group that is selected by virtue of being markedly deviant will, with or without special treatment, tend to regress toward the mean value of the total population. Skeptics also question how our present programs could possibly prove effective, citing the evidence on using punitive measures to change behavior. Ethical questions are also raised about the morality of denying a driver treatment that he needs simply to provide a control group in an evaluative study. Skeptics can also take issue with the underlying assumptions of driver improvement, namely, that driver failure is the major cause of traffic crashes, that removal of the bad drivers will correct the situation, and that punitive measures will lead to improvement in driving behavior. Available evidence challenges each of these assumptions.

There are others, both researchers and administrators, who would welcome sound evaluation of programs. Such evaluation requires rank and file cooperation from the agency involved. The research must not threaten the administrator or those engaged in implementing the program. The researcher can provide reassurance by taking the position that the program works and he is interested in determining which parts work best. If it is not feasible to establish a control group, the standard treatment can be compared with treatment

modifications. Control groups can frequently be established when programs are first being implemented and there is not enough treatment to go around.

Evaluations usually focus on driver attitudes and knowledge, traffic citations, and collisions. Attitude changes should be viewed with caution, since they frequently occur (as measured by the evaluation instruments) with no associated change in behavior. Usually citations are the basis for being assigned to a treatment group, and it appears that programs have had more effect on citations than on collisions. Yet highway safety is ultimately concerned with a reduction in the frequency and severity of collisions, so that collision data appear the most appropriate criteria for evaluating driver improvement programs.

Dr. Kaestner advocates the use of a more liberal level of statistical significance in evaluating highway safety programs, arguing that it would be worse to eliminate a program that is working than to adopt one that is not working, especially if all programs are evaluated at periodic intervals. He also advocates two-tailed tests of significance because of the level of knowledge in the field at this time. Programs could conceivably have effects opposite to what is anticipated. In addition, if drivers are randomly assigned to treatments, a check should be made to assure that the randomization was successful, that is, that the groups do not differ significantly on relevant variables. This check should occur before treatment is instituted.

The paucity of research using theoretical models is lamented by Dr. Kaestner, although that which has been conducted has not proved encouraging.

In summary, Dr. Kaestner advocates avoiding the two extremes, one of accepting outright the validity of driver improvement programs, and the other of rejecting outright the possibility of any program proving effective. In evaluating driver improvement programs, he underscores the importance of considering the threats to internal validity outlined by D. T. Campbell. In addition, special programs should be developed for different types of drivers. Finally, cost-effectiveness and cost-benefit analyses of programs should

be included in any evaluative effort.

Dr. Ross discusses evaluation in terms of changes in legal controls, and more specifically in relation to legislation aimed at controlling errant drivers. In contrast to driver improvement programs, changes in traffic law are relatively easy to evaluate because the overall goals of the changes are more readily identified and measured. However, it is usually more difficult to employ methods of random assignment where the law is concerned, since by definition the law should apply equally to everyone. Because of such limitations, Dr. Ross recommends the use of interrupted time-series analysis to determine whether a particular legislative change has had a significant impact. He refers to the Connecticut crackdown on speeding to illustrate his point, and, like Dr. Kaestner, uses D.T. Campbell's model of checking for plausible rival hypotheses to account for his findings. He concludes that because a number of rival hypotheses cannot be reasonably rejected in the Connecticut case, the data do not demonstrate that the drop in deaths in 1956 was caused by the crackdown.

After presenting several cases where before-after studies were not adequate to determine whether legislative changes had a true effect, Dr. Ross focuses on the British Road Safety Act of 1967. This legislation provided for roadside breath testing for alcohol, the results of which could be used as a basis for further more accurate testing. Drivers with blood alcohol levels of .08 percent or higher were deemed guilty of a crime, and the penalty included loss of license for a year.

Following enactment of the legislation there was a drop in highway fatalities, and the law was touted as a rousing success. Dr. Ross traveled to Britain and compiled the necessary data to determine whether the observed decrease was indeed a significant drop that could be attributed to the legislation or whether it was an insignificant fluctuation or an artifact attributable to other factors.

His careful analyses of relevant variables led him to the conclusion that the law had a genuine effect upon subsequent traffic fatalities, and that the effect was probably brought about by drivers separating their drinking and driving in time and space. The impact of the legislation was eventually di-

luted, and Dr. Ross offers further hypotheses for this change.

Driver improvement programs can be implemented gradually, especially where initially there are not enough trained personnel to meet the need. Such gradual introduction allows for the establishment of appropriate control groups. In contrast, legislative changes are initiated abruptly and control groups are usually out of the question. The interrupted time-series method described by Dr. Ross can provide a viable alternative to the traditional experimental-control group design, so that evaluation can often be made of programs that have frequently been considered beyond the expertise of the researcher.

Dr. Kaestner and Dr. Ross have illustrated procedures that can be put to use in the evaluation of highway safety programs even though the evaluator cannot establish a carefully controlled research design and manipulate the relevant variables. While methodology will never replace imaginative thinking on the part of the researcher, the methodology described by this symposium's speakers can provide useful guidelines to the evaluative researcher.

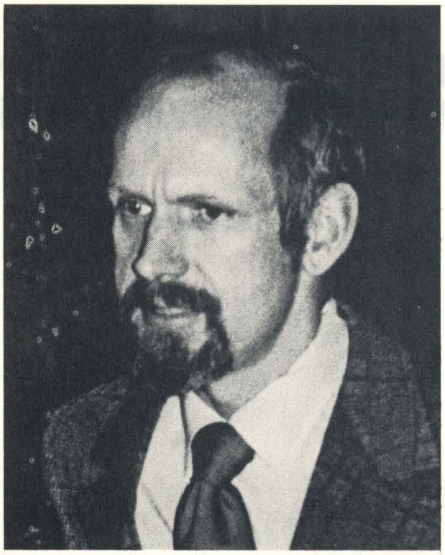
Evaluation of traffic safety programs is at best a complex undertaking. Problems of adequate records, vested interests, access to information (invasion of privacy), political feasibility, and inability to perform routinely accepted experimental manipulations lead many to seek other pursuits instead. Yet a growing social responsibility on the part of government makes it mandatory that programs be evaluated to determine how to gain the greatest benefit from the public tax dollar. Administrators and researchers must learn to communicate more effectively and to trust and respect the special competency of each other. Only in this way can we acquire the information so necessary for intelligent appropriation of effort.

Patricia F. Waller

Section I

The Impact of Driver Improvement: Do We Really Want to Know?

Noel F. Kaestner



DR. NOEL F. KAESTNER

Dr. Noel F. Kaestner is Professor of Psychology at Willamet University in Salem, Oregon. In addition, he holds a number of consultantships, one of which is with The Oregon Traffic Safety Commission.

In this latter capacity, Dr. Kaestner has produced some of the finest research in the field of highway safety. First focusing on highways but later examining the driver, his recent work has centered on driver improvement programs, an area of highway safety research in which he is a recognized authority.

THE IMPACT OF DRIVER IMPROVEMENT: DO WE REALLY WANT TO KNOW?

By Noel F. Kaestner

SCOPE AND PHILOSOPHY OF DRIVER IMPROVEMENT

Driver education refers to classroom and behind the wheel programs in the high schools for younger drivers who have not had extensive driving experience nor have been licensed motor vehicle operators. To a much lesser extent commercial driving schools play a role in driver education. By contrast driver improvement programs are directed at the changing sub-population of experienced and usually licensed drivers who, during a particular unit of time, have a significantly disproportionate number of collisions and/or citations for traffic law violations. Administration of the latter programs may be by law enforcement agencies, traffic courts, industry sponsored schools, or community service organizations -- the last ones often sponsor the National Safety Council's Defensive Driving Course. However, the most extensive driver improvement efforts are typically within the aegis of motor vehicle and driver licensing agencies, and furthermore the bulk of the evaluation effort has similar origins.

Driver education therefore is concerned with the preparation of initiate drivers for meeting minimal standards for the successful acquisition of a motor vehicle operator's license. Driver improvement programs are concerned with the maintenance of driving skills that will minimize collision involvement and traffic law violations.

Although driver improvement programs administered by motor vehicle and licensing agencies traditionally have followed the guidelines of the American Association of Motor Vehicle Administrators (1965), a relatively wide diversity of programs has emerged. Despite the idiosyncrasies of individual jurisdictions most programs incorporate several central components. Most typically a full scale driver improvement pro-

gram involves a three stage procedural format. Thus, the first contact by an agency with an errant driver typically involves a warning or advisory letter that attempts to encourage and/or threaten the driver with the object of motivating him to improve his driving performance. For drivers who do not respond favorably to the receipt of this mail contact, an interview or hearing is arranged for a face-to-face confrontation. These personal encounters may take the form of individual interviews or larger group meetings. Finally, those who do not adjust their driving habits to remove themselves from the watchful attention of the driver improvement officials have their driving privileges suspended or revoked. By far the most expensive component of this total package involves the second stage, interview or group meetings. The prominence of this stage in the overall program has been reflected in the disproportionate number of research studies that have been devoted to the evaluation of this phase.

THE ANSWER IS "NO" -- WE REALLY DON'T WANT TO KNOW

There is a wide spectrum of reasons why we may not want to know the impact of driver improvement measures as presently implemented. Historical precedent would support this position. The practice of using control groups in social science research is a relatively recent development. According to one researcher (Plutchik, 1968, p. 167), the first control groups appeared in the literature in 1908, and as late as 1933 only 11 percent of the studies in the literature employed control groups. Nevertheless, three and one-half decades ago researchers (Johnson and Cobb, 1938; Johnson, 1939) argued vehemently for definitive experimental designs embodying equivalent control groups to compensate for regression effects. Two decades later B. J. Campbell (1959) wrote with cosmic clarity on the limitations of ex post facto studies and strenuously urged the generation of controlled experimentation. Thus, there was a lag of nearly three decades before the first of the California, New Jersey, Oregon and Washington research efforts appeared -- a lag that poignantly reflects the sense of urgency about wanting to ask about the extent of driver improvement impact.

The reluctance to ask within a scientifically sound

framework about driver improvement programs in terms of the historical record is more descriptive than analytical. The underlying resistence to framing adequate research proposals is critically dependent upon cherished beliefs that have grown up with the "homemade" generation of individual jurisdictional programs. A factor contraindicating research is the "face validity" of the programs. The essence of the sometimes implacable dedication to existing programs has been embodied elsewhere in the statement that, "It is one of the most characteristic aspects of the present situation that specific reforms are advocated as though they were certain to be successful" (Campbell, D. T., 1969).

The ready acceptance of the face validity of existing programs has implications for not subjecting them to adequate research evaluation. Thus, a pseudo-ethical issue frequently emerges wherein arguments are made that it is unfair and immoral to assign errant drivers to control groups thereby denying them the advantages and benefits of the driver improvement measures. This concern was touched upon in the previously cited paper by B. J. Campbell, (1959). There he stated, "Objections may be voiced to 'experimenting with human lives'; but these objections can be met by proceeding carefully. No driver need be callously handled for the sake of experimentation. Even if such were necessary, it would seem that a clearer conscience would be justified than is the case now upon considering that many drivers die each year because such experimentation has not taken place and thereby has not resulted in the refinement of techniques to the degree of maximum ability to influence dangerous drivers."

A dramatic firsthand illustration of this ethical dilemma occurred in an unpublished pilot study in Oregon. In it a particular driver was assigned to the control group. He then had a traffic citation for a moving violation and was thereafter recorded as a failure in the control group. He was subsequently assigned to the traditional driver improvement interview that was under investigation in the study. After his interview he was involved in a culpable traffic collision which cost him his life. (The strategy adopted in the Oregon studies has been to assign members of the control group to the investigated treatment condition as soon as they incur either a moving violation or a chargeable traffic collision.)

Closely related to the face validity argument against control group evaluations of driver improvement efforts, is the construct validity interpretation. Influenced no doubt by the strong AAMVA emphasis on improvement of driver attitudes and the work of Tillman, et al., (1964) which stressed that a man "drives as he lives", evaluation efforts have often taken the form of before-after comparisons of driver attitudes. A review of the literature by this writer uncovered innumerable references to this type of evaluative study with the invariant findings that: (1) post-treatment attitudes typically changed significantly in a generally favorable direction; and (2) no communsurate change in driving behavior was noted or even measured. The acceptance of improved driver attitudes as revealed by paper and pencil assessment instruments mitigates against the generation of adequate research studies on the impact of driver improvement on the primarily non-verbal driving performance. Because of the basically non-verbal components of most driving skills, the construct validity assumption that whatever improves driver attitudes will inevitably improve actual driving performance must be rejected.

Probably one of the major roadblocks to the development of effective research designs for assessment of driver improvement programs consists of the inability to appreciate the regression toward the mean phenomenon. The most emphatic and energetic defense of current driver improvement actions, for example the hearing phase, consists of the argument that of the x number of drivers called in for the personal interview only some smaller fraction, x/k, does not show improved subsequent driving records and thus requires the follow-up suspension action. This argument, though simplistic, is tenaciously held by most motor vehicle division administrators and driver improvement program directors. They generally find it impossible to conceive of a situation wherein an even smaller than x/k fraction of drivers might receive suspension notices if they were <u>not</u> interviewed. Generally, psychologists and statisticians working with these agencies have not been universally effective in disabusing the administrators and program directors of the fallacious nature of this argument.

The forces described above that contrived to discourage objective evaluation derive from that segment of driver improvement officials who were generally convinced of the posi-

tive benefits of driver improvement programs as currently administered. Either because of the acceptance of face or construct validity of the programs or the ready acceptance of pre-post study results, they feel no need to embark upon research programs requiring additional technical competence. At the opposite end of the spectrum are those individuals who reject the need for driver improvement evaluation either because they feel its philosophical and pragmatic bases are unrealistic or because enough prior research has been generated to cast serious doubts on the contributions of driver improvement measures.

The philosophical underpinnings of contemporary driver improvement have been considered in greater detail elsewhere (Kaestner, 1972). There the four specific objectives outlined by the AAMVA (1965) were enumerated and the implications of each for driver improvement programs spelled out and criticized. The first objective involves improving driver attitudes and driving performance and instilling the will to improve. Objectives 2 and 3 are concerned with determining whether problem drivers are afflicted by physical or mental deficiencies and if so the application of appropriate restrictions. The fourth objective concerns itself with eliminating from the highways the unsafe, incompetent and physically or mentally unqualified drivers via refusal or withdrawal of driving privileges.

The enunciation of these principles bear evidence of the heavy commitment to certain principles. These are: (1) driver failure, particularly faulty attitudes, is a primary cause of traffic collisions; (2) removal of a few of the most recalcitrant drivers from the highways will drastically reduce the overall highway accident toll; and (3) punitive actions in the form of suspensions, revocations, etc., will effectively improve attitudes and instill the will to better driving practices.

To the extent that these assumptions are tenable the basic philosophy of current driver improvement is defensible. Considerable carefully researched evidence exists to question each of these principles, and consequently the fundamental philosophy of driver improvement. With regard to the first principle cited above there is no recognition of the approach

that views the driver as but one element in the man-machine-environment system. Furthermore, the central emphasis on driver attitudes and the will to improve denigrates the multiple causation approach to most traffic collisions so clearly elaborated by Blumenthal, (1968).

The second principle with its emphasis on a small core of errant drivers neglects the evidence that indicates that collision involvement of drivers is not a highly stable characteristic (Schulzinger, 1956; Coppin, McBride, and Peck, 1965). Thus driver improvement programs that attempt to zero in on an especially small recalcitrant segment of the population who during a short interval has disproportionately many collisions and/or traffic violations cannot be expected to have a dramatic impact on the overall traffic safety picture. With regard to the punitive response of driver improvement programs via license suspension or revocation, the complexity and frequently ineffective motivational value of punishment, documented elsewhere (Solomon, 1964), has generally been ignored. (A detailed accounting of this variable appears in Miller and Dimling [1969]). Because of the unrealistic basic objectives of the driver improvement guidelines as pronounced by the AAMVA, a number of researchers have come to dismiss driver improvement practices as unworthy of their research attention, and another cluster of evaluators have considerable misgivings about the ultimate capacity of driver improvement programs to have a significant impact on traffic safety statistics.

Some of the practical deficiencies of the current driver improvement programs are considered in greater detail elsewhere (Kaestner, 1972). In that paper some of the obvious discrepancies between industrial selection, training and retraining programs and present driver improvement practices are cited. Without covering these in detail, the absence of effective retraining measures and the concomitant need for problem diagnosis and effective instructors characterize driver improvement programs. The industrial analogy of requiring a man who has had several accidents on the job to stay home 30, 60, or 90 days, thus denying him the opportunity to practice or relearn the skill of his trade, and then return to the job with the expectation that he will somehow have regained a level of competence during this interval of idleness is a reflection

of current suspension programs for driver licensing.

A second point made in the earlier paper concerned the possibility that even with a questionable philosophy of driver improvement it would be possible for some driver improvement programs to be effective. A kind of Hawthorne effect (Roethlisberger and Dickson, 1946) could be operating. Supposed or actual benefits of driver improvement programs can derive from the official publicity and word-of-mouth communications about these programs. Studies in California (Coppin, Marsh, and Peck, 1965) and Washington (Toms, 1966) revealed that drivers invited to driver improvement meetings but not actually attending did as well as those who appeared for the meetings and definitely better than another group of uninvited control group drivers. Thus, it is possible for a wide variety of programs, however ill-conceived, to have an apparent effect on subsequent driving behavior when in fact it is the mere presence of the program that is responsible for any rehabilitative impact.

Outlined above have been some of the philosophical and pragmatic reasons for the understandable reluctance to continue the inquiry about the effectiveness of driver improvement programs. Above and beyond these considerations there has been an abundance of empirical evidence from adequately controlled studies that do not provide encouragement about the potential of driver improvement programs as significant forces in the reduction of traffic collisions. Without attempting to provide an exhaustive survey of the literature, a few of the studies that come to the writer's attention will be cited. The limitations of driver improvement programs will be documented for each stage of the traditional driver improvement sequence: namely, the warning letter, the personal hearing, and the suspension and/or revocation. Examples of negative findings will be considered in that order.

The impact of the warning letter was studied in Oregon and one of the findings was that there was no difference for either convictions for traffic violations or chargeable accidents between a control group that received no warning letter whatsoever and an equivalent group that received the standard form letter (Kaestner, Warmoth and Syring, 1967). More recently a California study showed that their standard warning

letter had no impact whatsoever on subsequent traffic collisions based on a comparison with a matched control group (Marsh, 1971).

As mentioned earlier, personal contacts with recalcitrant drivers in the second stage of driver improvement may be either on an individual basis or in terms of group meetings. In either case, ample evidence exists to cause one to question the impact of either procedure. With regard to individual contacts at least four studies bring into question the adequacy of the one-on-one confrontation as an effective driver improvement procedure. First there is an unpublished pilot study in Oregon that showed that those interviewed had slightly more, though not significantly so, collisions subsequent to the interview in comparison with a matched control group. Shortly after that study, a California study was published which showed that the individual interview had no impact on traffic collisions relative to a no treatment control group (Coppin, Peck, Lew, and Marsh, 1965). More recently the same state evaluated two separate individual interview procedures, one of which was the regular individual hearing (RIH) and the other involving an experimental individual hearing (EIH) specially designed by a psychologist who trained the driver improvement interviewers to conduct this revised hearing procedure (Marsh, 1971). Neither of these approaches to the individual interview had any accident reducing effects.

The group meeting approach to personal call-ins also has its share of negative findings. A series of studies by Henderson and Cole (1964, 1965, and 1966) showed that no accident reduction ensued participation in a group discussion procedure, and this finding is especially significant in view of the fact that individuals invited to this program were selected primarily because of their patterns of accident repeating. Coppin (1961) and Coppin, Marsh and Peck (1965) provided two more studies showing the deficiencies in group meetings in terms of subsequent accident reduction.

Although relatively little research has been done on the effectiveness of driver license suspensions or revocations, the data that exist provide little encouragement for supporting the concept of license suspension or revocation as a therapeutic device. Coppin and Van Oldenbeek (1965) found that 33

out of every 100 suspended negligent drivers and 68 out of every 100 revoked negligent drivers drove during the suspension or revocation periods as judged by conviction and accident records on file. Consistent with this finding is an Oregon study in preparation that shows that of the 44 per cent who responded to a questionnaire on driving during the suspension interval, 50 per cent freely admitted that they drove at least once during the suspension. Presumably any response bias in these data would operate in the direction of providing a conservative estimate of the actual prevalance of driving during the suspension. Preliminary data from this same study revealed that under some circumstances, particularly in the larger cities, suspension of the operator's license is as effective as no action whatsoever or may even be counterproductive.

So far, this paper has reviewed some of the reasons why interest in the evaluation of driver improvement has been less than enthusiastic. Reasons for not evaluating programs range from rejection because of the obvious face validity of the procedures to objections on theoretical and pragmatic grounds. Finally, there is the uncritical acceptance of research data that document the failures of traditional driver improvement programs at each stage from the warning letter to the license suspension or revocation.

Because of, rather than in spite of, the wide divergence of opinion of the need for continuing research on the effectiveness of driver improvement and the considerable prevalence of driver improvement programs throughout the states, there is still a need, whether universally recognized or not, to assess driver improvement activities. Such studies should not be concerned with their success or failure, but rather as exercises in identifying particular components of the programs that are productive and under what specific circumstances these measures are especially appropriate. In order to achieve this sort of evaluation effort, it is important to carefully plan and design experiments so some of the shortcomings and pitfalls of earlier studies may be avoided. To this end, the various threats to internal validity identified by Donald T. Campbell, (1969) will be enumerated and the special application of these threats to validity in the field of driver improvement will be examined. As a consequence of this kind of effort it may be possible to plan evaluation programs

that produce definitive results for program administrators and researchers alike.

THE ANSWER IS "YES" -- IF WE REALLY WANT TO KNOW, THEN ...

If we really want to know the impact of driver improvement, then it will be necessary to design research studies that recognize the various threats to the internal and external validity that Donald T. Campbell has so cogently outlined in his article entitled "Reforms as Experiments" (D. T. Campbell, 1969). These threats to internal validity are regarded by Campbell as rival hypotheses that might explain away an effect observed in quasi-experiments. In well designed true experiments the employment of adequate control groups and the principle of randomization at least theoretically minimize the possibility of the rival hypotheses explaining the effects or differences recorded in experiments.

Campbell, after enumerating the threats to internal validity that are listed below, reminds the reader that only "plausible" rival hypotheses need concern the researcher as potentially invalidating in laboratory experimentation. By contrast in correlation studies and so called "common sense" descriptive studies, more care is required to mitigate the very real possibility that the rival hypotheses may well be serious and not at all inconsequential alternative explanations for data trends occurring in the study. In the field of research on driver improvement, only occasionally does the study plan even approximate the true experiment, and by far the bulk of the evaluative efforts are quasi-experiments at best. It therefore seems worthwhile to devote some time here to consideration of these threats to internal validity individually. After enumerating these potential rival hypotheses, a series of examples of failure to control for these sources of confounding will be cited from the literature.

D. T. Campbell's Threats to Internal Validity

In the article cited above, nine threats to internal validity were presented. These were:

1. History: events, other than the experimental treat-

ment, occurring between pre-test and post-test and thus providing explanations of effects.

2. Maturation: processes within the respondents or observed social units producing changes as a function of the passage of time per se, such as growth, fatigue, secular trends, etc.

3. Instability: chronic unreliability of measures, fluctuations in sampling persons or components, autonomous instability of repeated or "equivalent" measures. (This is the only threat to which statistical tests of significance are relevant.)

4. Testing: the effect of taking a test upon the scores of a second testing. The effect of publication of a social indicator upon subsequent readings of that indicator.

5. Instrumentation: in which changes in the calibration of a measuring instrument or changes in the observers or scores used may produce changes in the obtained measurements.

6. Regression artifacts: pseudo-shifts occurring when persons or treatment units have been selected upon the basis of their extreme scores.

7. Selection: biases resulting from differential recruitment of comparison groups, producing different mean levels on the measure of effects.

8. Experimental mortality: the differential loss of respondents from comparison groups.

9. Selection-maturation interaction: selection biases resulting in differential rates of "maturation" or autonomous change.

In addition to the nine threats to internal validity, Campbell (1969) identifies five threats to external validity. As this term is used by Campbell, these threats may dilute generalizations from the original research setting to a wider range

of applications. Of these threats to external validity, the one that has the greatest relevance to research plans in driver improvement consists of "reactive effects of experimental arrangement." In this category Campbell includes the now classic "Hawthorne effect." We will illustrate how each of Campbell's threats to validity may affect evaluations of driver improvement programs.

History

Reductions in automotive fuel supplies and the imposition of mandatory or voluntary rationing constitute an impressive example of events, other than experimental treatments, that may occur between pre-test and post-test that provide alternative explanations of treatment effects. The research literature is replete with examples where the control and experimental groups were not run concurrently. In my review of the seven most prominently and fairly evaluated driver improvement programs, three of the seven did not collect interview and control data for the same intervals (Kaestner, 1968). The National Safety Council's research report entitled, An Evaluation of the National Safety Council's Defensive Driving Course in Selected States compares control and treatment groups data for non-overlapping intervals (Planek, et al., 1972). A recently completed ASAP three-year demonstration project in Oregon showed definite downward trends for average BAC (blood alcohol concentration) for arrested drivers and for drivers killed in traffic collisions (Mental Health Division, 1973). Interpretations of these findings is cloudy at best in view of the fact that during the course of the three-year project the legal definition of drunk driving was changed from .15 BAC to .10 BAC. Because it is very clear that the real world does not stand still during the course of research studies, it is strongly recommended that any effort that goes as far as the generation of control groups should extend the extra effort involved by insuring that control and experimental groups are assessed during identical real time intervals.

Maturation

The need for control groups is nowhere better illustrated than with regard to the factor of maturation. Maturation as used here refers to changes that might occur purely as a func-

tion of the passage of time. Most driver improvement programs focus on the younger driver. Median ages of drivers in the program at the letter, interview and suspension stages in three Oregon studies are 25, 23, 21, respectively. (This reversal of the expected trend for median ages results from the fact that these three studies extended over nearly a decade and the emphasis has shifted with regard to the target population in this interval.) These median age values are not too dissimilar from those occurring in California, New Jersey and Washington studies. Because of the youthfulness of this subpopulation of drivers, it is absolutely necessary to include matched controls of equivalent ages.

Instability

In other places D. T. Campbell has referred to this rival hypothesis as mere chance. Campbell points out that this is the only threat to internal validity with which statistical tests of significance are concerned. These tests typically compare some measure of variability ascribed to treatment effects with the inherent within treatment index of variability. Because the coefficient of variation (100x standard deviation/mean) for accidents is generally several times higher than that for traffic convictions, this accounts, at least in part, for the apparent greater impact of driver improvement programs on traffic convictions as contrasted to collisions. This discrepancy between the variability of convictions versus accidents was documented in an earlier California study (Coppin, McBride, and Peck, 1965). In correlating accidents with accidents and convictions with convictions over two successive time intervals, they calculated the coefficient of determination (r^2) and found that "although none of the obtained correlations are very high, the between-violations coefficients are several times higher than the between-accident coefficients. . . ." Although the topic of instability as well as all the other threats to internal validity are concerned with type I errors (concluding there is a real difference when there is none), it is conceivable that the extremely high variability of accidents relative to their own average, constitutes a real possibility of obtaining an inflated type II error (concluding there is no real difference when one actually exists).

With regard to instability in the application of statis-

tical tests, the procedures of the California group and the Oregon Research Institute involving the use of Mann-Whitney U tests are probably quite appropriate in view of the extreme skewness of distributions of traffic involvements. This caveat is especially noteworthy with regard to analysis of collision data.

Testing

Certain driver improvement programs, particularly local defensive driving courses or court-sponsored programs, offer as evidence of program success shifts in attitudes of attendees in favorable directions and/or increases in information level with regard to vehicle operation and traffic laws. These intermediate criteria are generated from pre- and post-testing of attitudes and knowledge. Unless comparable control groups that are similarly pre- and post-tested are employed, there is a real risk that the alternative hypothesis of the effect of taking the first test upon the subsequent score of the second test is responsible for the observed change. These comments are not meant to discourage the employment of intermediate measures, as the writer strongly favors the cause tree analysis (Driessen, 1970) or causal chain approach (Hall and O'Day, 1970). Certainly positive shifts in attitude and information level should not be disregarded. However, the need for controlling and testing rival hypotheses remains.

Instrumentation

Although calibration of measuring instruments is probably not a factor in the interpretation of findings of driver improvement studies, instrumentation problems have been interpreted to include what has been described in other places as experimenter bias (Rosenthal, 1963). Specifically we are referring to the fact that persons taking the measurements know the hypothesis and which group had the experimental treatment. This knowledge has a great potential for biasing the observations or judgments made in the research.

One of the chief applications of this principle would appear to pertain to the evaluation of traffic collisions and their classifications as culpable or non-culpable. It would appear that a "blind" procedure be employed so that the deci-

sion maker about culpability not be aware of the treatment condition of the driver whose traffic collision he is assessing. This requirement is an absolute must as the assumption of unquestioned experimenter honesty need not be violated in order that the experimenter bias be operative. It is suggested that the experimenter draw up guidelines for the decision making process prior to the examination of any accident records and that these decision rules be adhered to throughout the experiment. Furthermore, it appears reasonable that one decision rule would involve the employment of a third person judge, probably a driver improvement screening officer who is unaware of the treatment category of the operator, to decide culpability of collisions with especially ambiguous circumstances.

Regression

Probably none of the threats to internal validity is more serious than the regression toward the mean phenomenon. By the very nature of the driver improvement programs people are selected because of their extreme traffic records. Although some of the individuals certainly are on an upswing that might well continue without any official intervention, there is a wide variety of evidence that indicates that individuals accruing excessive numbers of citations or collisions during one time interval are not the same ones, for the most part, excessively involved in subsequent intervals.

Because of the universality of the statistical phenomenon of regression and the special appropriateness of this concept to the nature of drivers' records, it is absolutely essential to generate research designs that incorporate controls for this phenomenon. The most effective measure to accomplish this goal is the employment of equivalent control groups over concurrent time intervals. Failure to observe this rule irretrievably confounds program impact with spontaneous non-program driving record improvements.

Selection

This threat to internal validity simply states that unless two groups are equivalent along relevant variables at the beginning of a study, differences between groups receiving

different treatments at the end of the study cannot be definitively assigned to the treatment effects themselves. Thus, any biases among treatment groups serve as confounding variables and prevent unambiguous interpretation of the findings.

Within the driver improvement study four major variables have been identified as requiring attention in order to obviate selection effects. These are: driver age, sex, prior driver record, and traffic exposure. The California study (Coppin, 1961), a New Jersey study (Henderson and Kole, 1966), and a current Oregon study on suspension effectiveness all involve unfortunate matching of the age variable even though random assignment in treatment groups was involved. Both the California and New Jersey studies involved comparisons where the driver's sex was equated among the groups. (In the California study sex differences appeared in terms of percentages of drivers invited to the group interview and actually appearing which strongly favored the females.) With regard to the sex variable, the Oregon studies have generally been restricted to male drivers only on the premise that if the program components could be improved for male drivers, the program would have over 90 per cent effectiveness.

With regard to the equivalence of prior driver records, the second California group interview study (Coppin, Marsh, and Peck, 1965) and the New Jersey study cited above both involved inadequate matching. In the California case the prior driving records of the meeting group were milder than those of the control whereas the reverse situation obtained in the New Jersey study. In view of the fact that these were not true experiments as emphasized by Coppin in that other forms of Motor Vehicles Department intervention occurred during the post-treatment interval, it is difficult to use straight covariance measures to adjust for the inequalities in the groups in terms of prior driving record at the inception of the study.

As with the efforts to equate prior driving records through random processes, the randomization technique appears to be the only one that is generally employed to match samples on geographical distribution and the driving exposure variables. Studies that have come to the attention of the writer have not typically provided data on the effectiveness of the randomization procedure.

Experimental Mortality

Here we are referring to the fact that certain types of drivers may have dropped out of the study in a pattern which produces differences -- but the differences are due to differential dropout and not due to the treatments. Because of the generally transient nature of the driver improvement population it has been the policy in all Oregon studies to ascertain the resident status of each driver at the termination of the interval of record comparison. The rationale for this is the obvious need to establish whether a driver who has maintained a clear record during the term of the study has done so because of the absence of traffic entries on his record or because of his absence from the state and therefore from a reporting jurisdiction. A differential dropout rate was encountered in an Oregon interview study (Kaestner and Syring, 1967). The difference in mortality rate had the impact of favoring the interview group and was large enough to achieve statistical significance at the .05 level. No explanation for this differential dropout rate has ever been uncovered, but its impact despite its significance would not have changed the overall interpretations of the effects of the interview. Several years ago during the Vietnam engagement, differential mortality was a more considerable factor than is the case today.

Selection-Maturation Interaction

The only point to be made here is that if control and treatment groups are not equivalent on the age factor (selection) then the effects of the ordinary course of time and concomitant driving experience (maturation) may well have differential effects that make the identification of selection-maturation interaction inseparable from differences otherwise justifiably attributed to treatments. Again, because of the heavy emphasis of driver improvement programs on the very youthful driver, this selection-maturation interaction effect should not be regarded as a negligible rival hypothesis to explain obtained differences between groups.

PERSONAL OBSERVATIONS ON DRIVER IMPROVEMENT RESEARCH

In the following paragraphs some general considerations

about the initiation and implementation of evaluative studies of driver improvement will be outlined. For convenience, these have been set down under four sub-topics. The first of these concerns the researcher's role vis-a-vis program administrators--here referred to as "the foot in the door." Next there will be a brief reference to the choice of criterial measures of program success or failure. Logically following these remarks are some comments about certain statistical considerations that should be central in research design planning. This selection will close with a short discussion of the transfer of theoretical models in the area of driver improvement.

The Foot in the Door

For various reasons, some of which are nicely treated within Donald T. Campbell's article, employment of tactful strategies for the successful initiation of evaluative studies is an absolute must. The introduction of assessment procedures directed at more fully understanding the impact of existing programs may constitute a threat because of the considerable ego-involvement in their origins and implementations. Assessment can be threatening to the extent that defensiveness mounts to a point that efforts to evaluate the program are frontally negated or circumspectly made non-functional through subtle frustrations and absence of cooperation. My personal experience has been that it is best to propose evaluation schemes, not to determine if the program is working, but rather to more clearly demonstrate the full extent of the program effectiveness. In other words a useful working hypothesis is to assume the program is working and the thrust of the research project is merely to refine the measure of the efficacy of the various aspects of the total package.

Once the concept of program evaluation has been sold to administrators, various stumbling blocks still may exist. One of the prime sources of difficulty is the resistance often expressed to the establishment of control groups versus the standard treatment group. A more effective strategy might well be to compare the standard procedures with one or more variably modified treatment efforts. The researcher should encourage the administrators to contribute to the development of such program modifications. In this way, administrators

need never feel that they have "on their consciences" any disastrous consequences that may befall drivers assigned to the control group and thus denied the access to the treatment modality otherwise available.

Under some circumstances the generation of no contact control groups is both desirable and necessary. Even here program administrators' cooperation will be enhanced if the timing of the research proposal is propitious. For example, a higher level of administrator cooperation can be expected if the suggestion of assignment to a no treatment control group corresponds to a time when an especially heavy backlog of cases is prevalent. The introduction of control groups, randomly selected from the pool of potential driver improvement cases, will thus alleviate the heavy burden of an excessive number of waiting cases, and concomitantly provide the researcher with the generation of an adequate control group.

What to Evaluate?

Here we are concerned with the nature of adequate criterial measures. What signs will we go by to determine whether the driver improvement efforts have made a positive contribution to the traffic safety picture? In other words, what measures of program success are appropriate? The three most common measures of program impact have been changes in driver attitudes and knowledge, traffic citations, and traffic collisions. With regard to the first of these, I would urge that attitudinal and knowledge shifts be used as measures of program effectiveness that are supplemental to other ultimate measures. Various reviews of the literature have uncovered numerous instances where the desired attitude and knowledge shift reflected the highest aspirations and expectations of the program managers, only to find no parallel shift in either citation or collision experience.

With regard to citations and collisions as criteria, it would be best if we would heed the remarks of B. J. Campbell when he wrote, "It is basic that driver improvement programs need to contact the drivers most likely to cause or be involved in motor vehicle accidents and through appropriate action to reduce this tendency. At first glance it would seem the departments would therefore select for action those driv-

ers who have the most accidents. The actual and universal fact is, however, that action is initiated toward most drivers not because they have been involved in accidents, but because they have been convicted of violations of the motor vehicle law (most of which had nothing to do with an accident)" (B. J. Campbell, 1958, p. 13). This paradox continues to be as true today as it was in 1958. Thus, if drivers are brought into driver improvement programs because of excessive citations, then it would appear that program success be more appropriately measured in terms of subsequent citation data. By contrast, if the program is concerned with accident reduction, then the clientele should be selected on the basis of this aberation on their driving records, and subsequent traffic collisions become the appropriate criterial measure.

Because of the bias in the selection process in favor of traffic citations and the apparent differential effect of many of the programs in the direction of subsequent reductions in citations with little, no, or negative effects on traffic collisions, I am inclined to regard traffic citations as something less than the ultimate measure of program performance. Collision data, delays to collisions, culpability of collisions, extent of property damage or personal injury, etc., are some of the ultimate indices of program impact. The preference for this indicator obviously has implications for the study design in that longer time intervals and larger samples will probably be required. This may well be inconvenient in certain circumstances but certainly becomes a necessity in evaluating the impact of programs that have the stated goal of making our highways safer.

Statistical Considerations

In this section three areas that have special significance for research in driver improvement will be considered. These concern: (1) choice of significance level (alpha), (2) one-tailed versus two-tailed tests, and (3) post-randomization equivalence checks.

With regard to the alpha levels of statistical significance the researcher embarking upon studies in this area should seriously consider employment of alpha levels higher than those traditionally applied in laboratory research. Levels of .10

and .20 provide the greater sensitivity desirable at the exploratory levels of investigation. This point of view has been cogently expressed as follows by Marsh (1971) when he explained the use of an alpha level of .20. "This choice was made under the assumption that it would do more harm to reject a truly effective program than to adopt an ineffective one. This is also justifiable under the assumption that any program which is adopted can be (and should be) subjected to periodic evaluation, but that once a program is rejected it probably will never be re-evaluated."

Not at all independent of the recommendation to employ higher alpha levels is the argument that all statistical tests of significance of differences between treatment and control groups be performed as two-tailed tests. Several studies cited in the earlier portions of this paper revealed that on certain performance indicators, notably traffic collisions, the records of control group drivers were better than those of the drivers in the treatment condition. Especially in view of the differential impact of driver improvement programs on traffic citations versus collisions it seems especially appropriate to leave the door open for possible assessment of differences opposite to those anticipated for effective driver improvement programs. In other words, should a particular driver improvement measure be characterized by more collisions than found among no-contact control group drivers, it is necessary to be able to identify this discrepancy so that greater scrutiny to such a program will be forthcoming.

The final point here concerns the effectiveness of randomization in generating truly equivalent pre-treatment control and experimental groups. Again, a considerable number of studies have employed randomization with the goal of equalizing comparison groups only to find that the randomization process has failed miserably to equate the groups on such significant variables as age, sex, previous driving records, etc. In circumstances where analysis of covariance techniques is not appropriate, I strongly urge that the effectiveness of randomization be examined prior to the introduction of driver improvement intervention measures.

Obviously, by the very nature of randomization, the equivalence of comparison groups cannot be strictly guaranteed.

Nevertheless, the inequality of comparison groups appears to emerge far more often than chance expectations would predict in this field of driver improvement. I suspect that the strict rules for the operational implementation of randomization procedures lose something in translation from the researcher in charge of the program as they filter down to the clerical levels that are the terminal choice points in case assignments. The greater the distance between the scientist and the clerical helpers, the greater the likelihood randomization will fail in its mission.

Theoretical Models

For reasons not at all fully understood by this writer, the direct application of theoretical models of behavior to driver improvement settings has been somewhat less than a complete success. The implementation of a thoroughly thought through and pretested group dynamics model in the State of Washington failed to produce the desired results (Toms, 1966). This writer had the opportunity to participate in this program as a ringer, i.e., he was provided with a fake driving record with numerous entries and a call-in letter. On the surface, the group interaction and the occasional insightful remarks of participants seemed to be a hopeful sign that the model was working. However, other indicators from the behavior of the participants during the break midway in the session belied the earlier behavior as quite superficial.

Still another State of Washington study employed Skinnerian principles of behavior modifications with problem drivers (Kleinknecht, 1969). This program, though carefully planned and coordinated, also fell considerably short of expectations with regard to rehabilitative impact. In retrospect, the reasons for the incomplete success of this program are somewhat clearer. Very generally for behavior modification procedures to be effective, the experimenter, educator, or therapist must be able to control the significant contingencies. It is very likely that a great many potent reinforcers of careless driving are operative which offset the strictly manipulated rewards under the control of Kleinknecht.

A California study (Marsh, 1971) compared five group meeting techniques and two individual hearings with no con-

tact control groups. Two of the five group meeting procedures, the Subject Interaction Meeting (SIM) and the Leader Interaction Meeting (LIM), and one of the individual hearings, the Experimental Individual Hearing (EIH), were especially designed by a psychologist with many years of clinical experience. Not only did the psychologist develop the techniques that incorporated the then known most effective group procedures, but also he personally supervised the extensive training of eight special driver improvement analysts that were the only ones involved in the group and individual meetings. During the same time span the DMV staff developed two group meeting procedures, the Group Educational Meeting (GEM) and the Group Administrative Review (GAR). In addition to these there was a standard group program, a Driver Improvement Meeting (DIM) and the Regular Individual Hearing (RIH). The results in regard to traffic collisions revealed that the two group hearing techniques designed by Motor Vehicles Division staff, the GEM and the GAR, were significantly better and significantly worse, respectively, than the control group. The three methods designed by the clinical psychologist, reflecting the best in contemporary interpersonal dynamics, all involved more collisions than the control group and one of them, the LIM, nearly significantly so.

The failure of the "psychologically sound" procedures to reduce collision level below that of the control group might be explained by the fact that the clinicians did not conduct the hearings and that the effectiveness of a basically sound technique was lost when non-clinical personnel with relatively brief training administered these meetings. If this reasoning is followed, it is then difficult to account for the fact that the DMV-developed group hearing administered by even less well trained program people resulted in a significantly fewer number of collisions than the control group.

CONCLUSIONS

In summary, much of the evaluative research in driver improvement has not been impressive. On the one hand, there has been considerable naivete, encouraged by historical precedent, with regard to the concepts of face validity, construct validity, and regression effects. In quarters where naivete has

not been rampant, scepticism has been prevalent, often founded on philosophical, pragmatic, or empirical bases. What appears to be needed is the re-dedication to the principle that anything worth doing is worth doing well, and by this is meant the principle should be applied to evaluation procedures. Otherwise, there is no way of knowing whether the program itself is worth doing at all. To this end researchers are encouraged to re-examine the threats to internal validity expressed by D. T. Campbell and consider them in the light of the particular contingencies of the driver improvement setting. This re-dedication should also involve special consideration of administrative, criterial, statistical, and theoretical model problems.

As expressed elsewhere (Kaestner, 1968) there is a continuing need to develop diagnostic devices in order to provide tailor-made programs to the needs of individual problem drivers. The California studies by McBride and Peck (1969) and Marsh (1971) are certainly steps in the right direction. There is also a need to develop a broader base of clientele so that incipient problem drivers are recipients of program attention earlier in their driving careers.

Finally, more attention should be paid to cost effectiveness and cost benefit analysis of programs that do have significant impacts on traffic collisions. The two California studies just cited are noteworthy in this respect. An unpublished Oregon report showed similar savings to the public also. It is important that these analyses not be restricted to departmental savings at the operational level, but that they also incorporate savings to the drivers in the state from accidents that did not occur. In conclusion, driver improvement research today is at a point that corresponds to a state of affairs described by Beethoven paraphrased to the effect that the solutions to important problems are not so much to be discovered as to be invented.

REFERENCES

American Association of Motor Vehicles Administrators. Policies and position statements. Washington, D.C.: Author, 1965.

Blumenthal, M. Dimensions of the traffic safety problem. Traffic Safety Research Review, 1968, 12, 7-12.

Campbell, B.J. (Institute of Government, University of North Carolina), Driver improvement: the point system. Chapel Hill, North Carolina: American Association of Motor Vehicles Administrators, 1958.

Campbell, D.T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Coppin, R.S. A controlled evaluation of group driver improvement meetings. Sacramento: California Department of Motor Vehicles, 1961.

Coppin, R.S., Marsh, W.C., & Peck, R.C. A re-evaluation of group driver improvement meetings. Sacramento: California Department of Motor Vehicles, 1965.

Coppin, R.S., McBride, R.S., & Peck, R.C. The 1964 California driver record study - Part 6 - The stability of reported accidents and citations. Sacramento: California Department of Motor Vehicles, 1965.

Coppin, R.S., Peck, R.C., Lew, A., & Marsh, W.C. The effectiveness of short individual driver improvement sessions. Sacramento: California Department of Motor Vehicles, 1965.

Coppin, R.S., & Van Oldenbeek, G. Driving under suspension and revocation. Sacramento: California Department of Motor Vehicles, 1965.

Driessen, G.J. Cause tree analysis: measuring how accidents happen and the probabilities of their causes. Paper presented at the 78th annual meeting of the American Psychological Association, Miami, September 1970.

Hall, W.K., & O'Day, J. Causal chain approaches to the evaluation of highway safety countermeasures. Paper presented at the 37th annual O.R.S.A. meeting, Washington, D.C., April, 1970.

Henderson, H.L., & Kole T. Mass communication and group discussion techniques. New York: Drivers Safety Service, Inc., 1964.

Henderson, H.L., & Kole, T. Effective leadership through group discussion. Traffic Digest and Review, 1965, 13, 4-7.

Henderson, H.L., & Kole, T. Driver improvement clinics of the state of New Jersey. New York: Drivers Safety Service, Inc., 1966.

Johnson, H.M. Evidence for educational value in drivers' clinics. Psychological Bulletin, 1939, 36, 674-675.

Johnson, H.M., & Cobb, P.W. The educational value of drivers' clinics. Psychological Bulletin, 1938, 35, 758-766.

Kaestner, N.F. The 'state of the art' of research in driver improvement. Traffic Quarterly, 1968, 22, 497-520.

Kaestner, N.F. Human factors in driver improvement. In T. Forbes (Ed.), Human Factors and Highway Safety. New York: Wiley and Sons, 1972.

Kaestner, N.F., & Syring, E.M. Accident and violation reduction through brief driver improvement interviews. Traffic Safety Research Review, 1967, 11, 99-105.

Kaestner, N.F., Warmoth, E.J., & Syring, E.M. Oregon study of advisory letters - the effectiveness of warning letters in driver improvement. Traffic Safety Research Review, 1967, 11, 67-72.

Kleinknecht, R.A. Behavior modification of problem drivers. Olympia: State of Washington, 1969.

Marsh, W.C. Modifying negligent driving behavior: Evaluation of selected driver improvement techniques. Sacramento: California Department of Motor Vehicles, 1971.

Miller, L., & Dimling, J.A. Driver licensing and performance. Vol. 1. Research review and recommendations. Washington, D.C.: U.S. Department of Transportation, 1969.

Oregon Mental Health Division. Oregon alcohol safety action program - final report. Salem: State of Oregon, 1973.

Planek, T.W., Schupack, S.A., & Fowler, R.C. An evaluation of the National Safety Council's defensive driving course in selected states. Chicago: National Safety Council, 1972.

Plutchik, R. Foundations of experimental research. New York: Harper and Row, 1968.

Roethlisberger, F.J., & Dickson, W.J. Management and the worker. Cambridge: Harvard University Press, 1939.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton, 1966.

Schulzinger, M.S. The accident syndrome. Springfield, Illinois: Charles C. Thomas, 1956.

Solomon, R.L. Punishment. American Psychologist, 1964, 19, 239-253.

Tillman, W.A., et al. Group therapy amongst persons involved in frequent automobile accidents. Washington, D.C.: U.S. Army Medical Research and Development Command, 1964.

Toms, D. Pierce county pilot-study. Paper presented at an A.A.M.V.A. sponsored conference and workshop, Sacramento, California, April 1966.

Section II

Interrupted Time-Series Methods for the Evaluation of Traffic Law Reforms

H. Laurence Ross

DR. H. LAURENCE ROSS

Dr. H. Laurence Ross is Professor of Sociology and Law at the University of Denver. His wide range of interests includes the social impact of legislation.

In the area of traffic safety, he has examined the effects of the 1955 Connecticut crackdown on speeders, and, more recently, the impact of the British Road Safety Act of 1967, designed to reach the drinking driver. By bringing sophisticated methodology to bear on such problems, Dr. Ross has been able to introduce scientific evidence into areas previously dominated by opinion and political rhetoric.

INTERRUPTED TIME-SERIES METHODS FOR THE EVALUATION OF TRAFFIC LAW REFORMS

By H. Laurence Ross

INTRODUCTION

It may be true that we are entering the "experimenting society"--the welcome age when public administrators formulate policy on the basis of trial and evaluation. In this hoped-for, forthcoming, and perhaps even imminent stage of history, the administrators' commitments will be to problems rather than to solutions, and the public will expect imagination and honesty rather than consistent success from its policymakers. In this society, I conceive a principal role of the social scientist to be the provision for administrators of the tools whereby they can obtain competent evaluation of their programs; for it is only when the tools of evaluation are available that we can, in our roles as citizens, encourage the trial of new and even radical solutions to persistent social problems. In the absence of evaluation we run the risk that the political commitment needed to enact costly programs will require the continuation of well-intentioned but marginal, worthless, or even harmful activities because their costs and benefits are unknown and their abandonment would be politically embarrassing.

The scientific foundations of the experimental society are currently being laid by the development of formal analysis of experimental and quasi-experimental designs for research, and the elaboration of new methodological models for use in various real-world settings where important questions of cause and effect are being raised. I wish to present here some of the fruits of this effort in the matter of studies of law generally, traffic law in particular, and the drinking driver as an example.

In brief, my thesis is that studies of legal policies encounter inherent barriers to the use of the most satisfactory

of experimental designs. These difficulties arise from both practical and ethical considerations. Although in traffic law they may be more manageable than elsewhere--and good classical research designs have been implemented with success in some traffic law studies--in most legal impact studies resort must be had to quasi-experimental designs, which do not assume the ability of the researcher to assign treatments to subjects at random. Among these latter designs, the interrupted time series seems especially well suited to matters like traffic law, where there is available a valid measure relevant to the desired effect of legal controls, and this measure is routinely compiled over extended periods of time.

In this paper I will discuss the logic of the interrupted time-series experiment, illustrating the problems it is designed to overcome and referencing some technical materials now available for the use of this research design. The principle illustration of positive results utilizing interrupted time-series analysis will be my study of the British Road Safety Act of 1967, but for comparative and illustrative purposes I shall discuss as well several other time-series studies of the legal control of the drinking driver, and additional studies on the more general topic of evaluating legal efforts to control behavior, both in traffic and other problem areas.

* * * * *

In an experimenting society, the commitment of the public administrator to any legal policy will be a limited one, contingent on obtaining adequate evidence of the policy's effectiveness in achieving the administrator's goals at a cost deemed reasonable in relation to the value of the goals achieved (D.T. Campbell, 1969). Evaluation will be demanded for all major legal undertakings, and laws will therefore be implemented in a way that enhances the chances for rigorous evaluation of their effect. Policies that appear to be ineffective, or not effective enough to justify their cost, will be abandoned. However, where these policies relate to impera-

tive social needs, their abandonment will not signify political desperation, but rather a call on the administrator to try alternative policies.

Life today does not generally satisfy this model. Political commitments necessary to obtain legislation interfere with the ability to abandon a policy when it is shown not to be worthwhile. Worse yet, this commitment operates to avoid evaluation, or to skew evaluation so as to indicate success when little or nothing is being achieved. The situation is exacerbated where organizations are set up to implement a policy and the members feel that their jobs are at stake in maintaining it. The traffic law area has certainly been one in which political commitment and needs for organizational survival have impeded evaluation, and we have recently been reminded by a prestigious scholar that most of the common institutions used to deal with the matter of traffic safety are based on little or no evidence of effectiveness (B.J. Campbell, 1970). These include driver licensing examinations, driver education, motor vehicle inspections, and many other taken-for-granted programs.

There are, however, some signs that the situation is changing. Most major social welfare programs of recent national administrations have been accompanied by competent, if controversial, attempts at evaluation, and in some cases the administrators involved appear to have paid attention to the results. Even in the dim corner of public policy devoted to traffic safety, some light is appearing. Significant research has been funded by the Federal government through the National Highway Traffic Safety Administration and by private foundations such as the Insurance Institute for Highway Safety. Although we may be far from an experimenting society in the sense of one where all continuing programs are based on demonstrations of effectiveness, there is currently an interest in evaluation on the part of many administrators, and social science is being challenged to produce methodologies to respond to this interest.

Traffic law is in principle one of the easier fields in which to produce evaluation because its goals are relatively clear in comparison with other fields of law. One might question, for example, whether the goals of a taxation policy were

those of raising money, of discouraging some kind of private consumption, or of obtaining adequate records concerning the circulation of some sensitive commodity. Is the goal of school integration a matter of securing equal academic performance from both blacks and whites, preventing members of the minority from acquiring crippled egos, or establishing an abstract principle of justice which is fulfilled by the mere fact of integration itself? These questions are certainly arguable. In the area of traffic control and regulation, however, the goal of reducing accidents, injuries, and deaths occupies an uncontested first place, with efficiency of travel being an additional consideration. Furthermore, as social phenomena go, traffic-related fatalities are fairly well measured and give a relatively objective measure of the most fundamental goal to be achieved by traffic law. (However, fatal accidents are relatively rare and provide an unstable index for evaluating programs over short times and in small jurisdictions, and many studies are thus forced to rely on the far less reliable general accident statistics; see Zylman, 1972.)

The traditional answer of social science to the question of how to secure the most valid cause-and-effect knowledge has been to suggest randomized, comparative, classical experimental designs (D.T. Campbell & Stanley, 1963). This is still the best answer to the question. However, the defining characteristic of randomized, comparative experimental designs is that they involve the allocation of treatments at random to experimental and control populations. This random allocation is often difficult to achieve in studies on non-laboratory populations, and I suggest that studies in legal effectiveness are particularly difficult to undertake in the classical experimental mode.

There are three sets of reasons for this fact. First, from the practical perspective, it is difficult to formulate laws in such a way that they apply differently to similarly-situated individuals, which is exactly what is required by experimental designs. Second, there are peculiar ethical problems in legal studies that are not met elsewhere. Law embodies the principle of equal treatment, which may be interpreted to be precisely in opposition to comparative experimental assignments. Moreover, if a legal consequence can be

regarded as penal, it involves the principle that the punishment should fit the crime and perhaps also the criminal, whereas experimental treatments may well involve some conditions where the penalty seems at first glance to be inappropriate to both the one and the other. Even if the researcher finds that his specific design passes his own ethical muster, he may have to convince cooperating legal actors of the validity of his position. Third, designs which avoid the above-mentioned problems may well be difficult to implement because legal actors are strongly integrated into a system of mutual obligations, e.g., between judges and attorneys, where mutual favors may prove more pressing than promises to outside researchers (Ross & Blumenthal, 1974). These considerations do not completely preclude experiments on legal impact, but experimental designs in this area are understandably rare, and the investigators frequently report the emergence of unanticipated problems.

The class of research designs known as quasi-experimental has in common the acknowledgment of limitations on the ability of researchers to exercise control over the application of treatments. Because these limitations are frequent in practice--as in the study of law--quasi-experimental designs are attractive techniques for evaluation research, despite inherent weaknesses in comparison to classical experimental designs. Among these designs is the interrupted time series, which I propose as a standard method applicable in a wide variety of studies of legal impact, especially in the realm of traffic law.

The reasoning underlying interrupted time series is that if a cause-and-effect relationship exists between two variables, a discrete change in a causal variable will be reflected in an appropriately timed change in the effect variable. This is the same logic on which the much more common but unsatisfactory before-and-after study is based, but the time series is capable of yielding much firmer knowledge. To illustrate the interrupted time series in contrast to the before-and-after model, and to demonstrate its advantages in leading to clear and interpretable conclusions, I will utilize the example of the 1955 Connecticut crackdown on speeders, which Donald Campbell and I studied some years ago (D.T. Campbell & Ross, 1968).

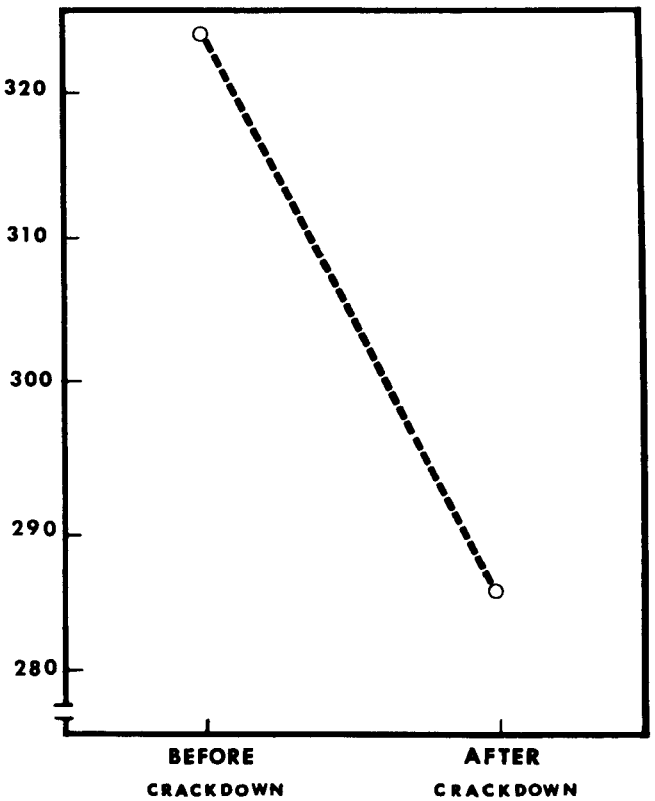


FIGURE 1. Connecticut traffic fatalities, 1955-56, as a before and after study.

Source: Campbell & Ross, 1968.

Most readers of this report will recall this particular reform, initiated by former Connecticut Governor Abraham Ribicoff. In 1955, a total of 324 people had been killed on the highways of Connecticut, and the public was demanding that something be done about the problem. On December 23 the Governor responded by the decree that henceforth all persons convicted of speeding would have their licenses suspended for thirty days on the first offense and for longer periods on subsequent offenses. The decree was enforced by threatening not to reappoint judges who refused to comply with it.

The months following Ribicoff's decree in general showed lower numbers of deaths than in the previous year, and the total deaths for 1956 were only 284, forty fewer than in 1955. The data are diagrammed in Figure 1, and seem to offer convincing evidence for the claims by the Governor that his reform had succeeded. However, it is evident to the methodologically sophisticated critic that the decline diagrammed here might well be explained by several groups of alternative causes potentially operating in the situation, which are known through past experience to produce results similar to those observed here. In the language of quasi-experimentation, these alternative causes are termed rival hypotheses. They must be eliminated as being implausible before we can accept the hypothesis that the experimental treatment or change in the putative causal variable in fact was responsible for the change in the effect variable. Among the many categories of rival hypotheses known, six are of primary concern in a situation like the one at hand. These are as follows:

1. History

The term refers to specific events other than the experimental treatment (the legal crackdown), occurring at about the same time, which might independently have caused the difference between the before and after observations. In this instance, 1956 might have been a year of bad weather, and we know from other evidence that in rain and snow, though accidents in general increase, serious and fatal accidents are less common. Again, for example, there might have been a significant improvement of 1956- and 1957-model cars, diminishing the chances that a passenger in an accident would be injured.

2. Maturation

The term originates in psychological studies, where it refers to changes related to the passage of time, such as growing older, tired, bored, sophisticated, etc. As opposed to history, which refers to discrete events, maturation refers to regular processes, the causes of which need not be understood. In this category falls the possibility of a general long-term trend towards declining accidental deaths, presumably due to better roads, medical care, etc. Were the Connecticut data in the form of a rate rather than absolute numbers of deaths, maturation would be a troubling rival hypothesis, for there was in fact such a decline in the mileage-based death rate in the United States for several decades including the period being discussed. The before-and-after study is insensitive to the possibility of such an explanation.

3. Testing

It is frequently found in science that measuring something will itself produce a change in the thing being measured, regardless of any experimental treatment intervening between the measurements. We know in the present instance that the 1955 death figure was considered extreme and disturbing, and it is possible that this fact alone produced safer driving and a lower number of deaths in 1956.

4. Instrumentation

Sometimes a before-after change will reflect a difference in the method used to measure a phenomenon, even though the phenomenon is actually unchanged. For instance, the weight of an object may appear to change when the scale being used becomes rusty. When dealing with social data, a shift in the way in which events are recorded can produce apparent changes, and such shifts often take place as part of general reform efforts that are centered elsewhere. One example (Sween & D.T. Campbell, 1965) is the enormous increase observed in the crime rates of Chicago in 1969 when a reform police administration was installed, presumably due to more complete reporting of crimes. Although there is no evidence of a record-keeping reform in the Connecticut case, the problem is potentially present in the desire of the administration to see its efforts

confirmed statistically.

5. Instability

An always plausible rival hypothesis is that the change is due to the instability of the measures involved. Whether a change in the effect variable is to be considered meaningful or insignificant depends on whether changes of its magnitude are routine or rare when purportedly causal events are not around. In the present case, if changes on the order of 40 deaths per year, up or down, were common in Connecticut, one would be inclined to question the meaningfulness of the observed change. If the rate were otherwise quite stable, rarely varying more than, say, 10 deaths per year in either direction, we would be much more impressed. The inherent instability of a particular social index is a function of a variety of matters including the population size on which the rate is computed, the reliability of measurement, and the number and strength of causal factors impacting upon the phenomenon.

6. Regression

If the before observation in a before-and-after study can be considered extreme, one can predict with some confidence that the after observation will be less extreme. In technical terms this phenomenon is known as regression toward the mean. Unfortunately, many policy changes are instituted exactly because a social problem has reached an extreme level, and evaluation of the effects of these policies is rendered problematic in a before-and-after study because of the plausibility of regression. The present case is unfortunately typical, the crackdown being instituted after a year of record-level deaths. The possibility of regression is a particularly bothersome rival hypothesis.

Because none of these rival hypotheses is ruled out by the before-and-after design, we must conclude that Figure 1 does not demonstrate any effect for the Connecticut crackdown. However, as with traffic studies generally, more data are easily available in this case. To switch to the interrupted time-series method, it is necessary merely to add statistics

of deaths in Connecticut for additional years before and after the initiation of the crackdown. The additional data enable us to rule out as plausible rival hypotheses all of the above-mentioned categories except that of history. This is so because, if our cause-and-effect understanding is true, we can expect a particular form of curve of the type illustrated in Figure 2A, which would not be produced by the factors named in the rival hypotheses. The figure shows a distinct break at the time of the experimental treatment, not continuous with prior or subsequent trends. Factors such as maturation, testing, and instrumentation are likely to be present at all points of measurement, and it is unreasonable to expect them to produce a change in the effect variable suddenly and coincident with the time of the treatment. These factors would more likely produce a curve like Figure 2F or H. Instability is observable in the routine differences from point to point, and it can be ruled out of consideration if the change at the point of introduction of the theoretical cause is deemed unusual by a test of statistical significance. Figure 2G shows a change plausibly explained by instability. Regression can be ruled out if the point prior to the reform is seen to be typical and not extreme. Even with an ideal form of curve, however, it is still possible that other causes occurring simultaneously with the theoretical one produced the apparent effect, and with a simple interrupted time-series design the researcher must shoulder the burden of raising all plausible alternative causes of a historical nature and ruling them out of consideration by external evidence.

Our exemplary case is transformed into a time series in Figure 3, which unfortunately lacks much resemblance to the ideally interpretable line of Figure 2A. It does seem that maturation, testing, and instrumentation are effectively ruled out, but the line demonstrates a great deal of instability, and a test of statistical significance does not permit rejection of the null hypothesis of no effect for the crackdown (Glass, 1968; also Glass, Willson, & Gottman, 1972). Moreover, the plausibility of an explanation in terms of regression is obviously very high. It cannot be said with any confidence that the decline in traffic deaths in Connecticut in 1956 was caused by Governor Ribicoff's crackdown on speeders.

Another example of negative findings concerning speed

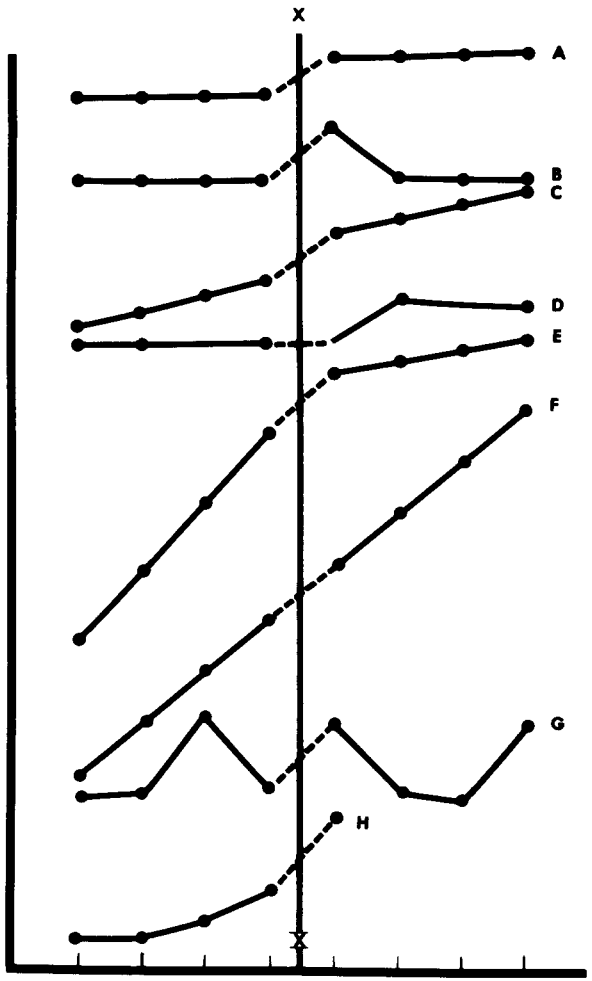


FIGURE 2. Idealized interrupted time-series data

Source: Campbell & Stanley, 1963.

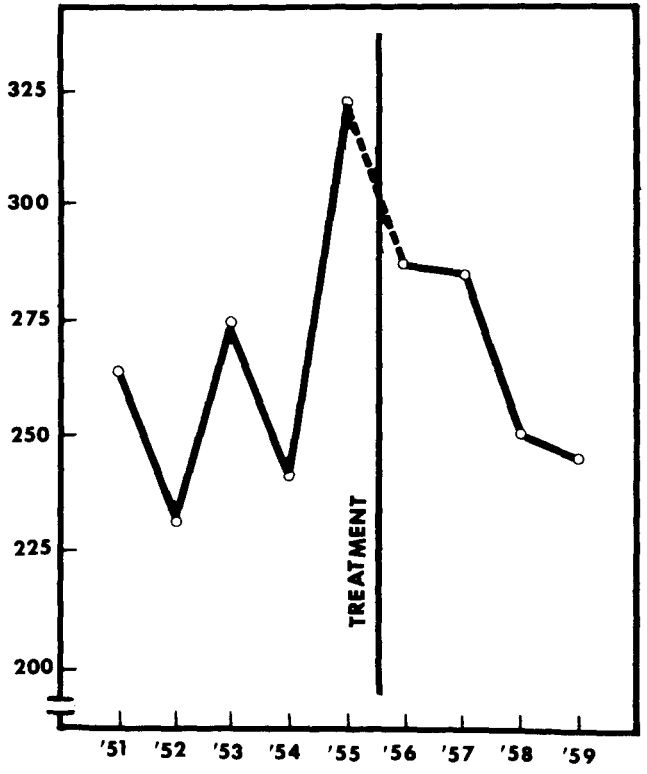


FIGURE 3. Connecticut traffic fatalities, 1951-59, as an interrupted time-series study.

Source: Campbell & Ross, 1968.

limitations and traffic safety is provided in Figure 4, based on a report wherein the Insurance Institute for Highway Safety (1973) criticized as premature claims that speed reductions induced by the energy shortage of that year reduced highway fatalities. These claims were being made on the basis of holiday weekend fatality data, which are available for analysis much more quickly than are complete fatality data. A sharp decline in deaths in selected states over the Thanksgiving holiday in 1973 as compared with 1972 was being interpreted as evidence for the safety effect of lower speeds. The I.I.H.S. figure presents a short time series putting the admitted decline in deaths into perspective, and revealing that these data alone do not support the claims in question.

Figure 5 shows how time-series analysis can produce visually convincing positive evidence of effectiveness for a traffic law. The data on motorcycle deaths in Michigan (Klein & Waller, 1970) clearly show the effect of two legal changes--the enactment and subsequent repeal of a law requiring motorcyclists to wear safety helmets. Note how this simple curve, based upon data easily accessible to the public, proves the effectiveness of the legislation with simple clarity. The plausibility of history as a rival hypothesis is strongly reduced by the response of the curve to both the inception and the termination of the law, and all other categories of rival hypotheses mentioned above are rendered implausible as discussed previously.

The utility of interrupted time-series analysis is well demonstrated in the study of laws aimed at controlling the drinking driver. In the following pages I will review the results of my own study of the British Road Safety Act of 1967 (Ross, 1973) and of several other studies of similar legislation in the United States.

Apart from Scandinavian laws, concerning which there is virtually no research, the British Road Safety Act has received the greatest amount of attention of all attempts to control the drinking driver through law. In brief, the Act provides that under certain circumstances, including involvement in a traffic violation or an accident, a driver may be required by a policeman to take a roadside screening breath test for the presence of alcohol in his body, and if this test

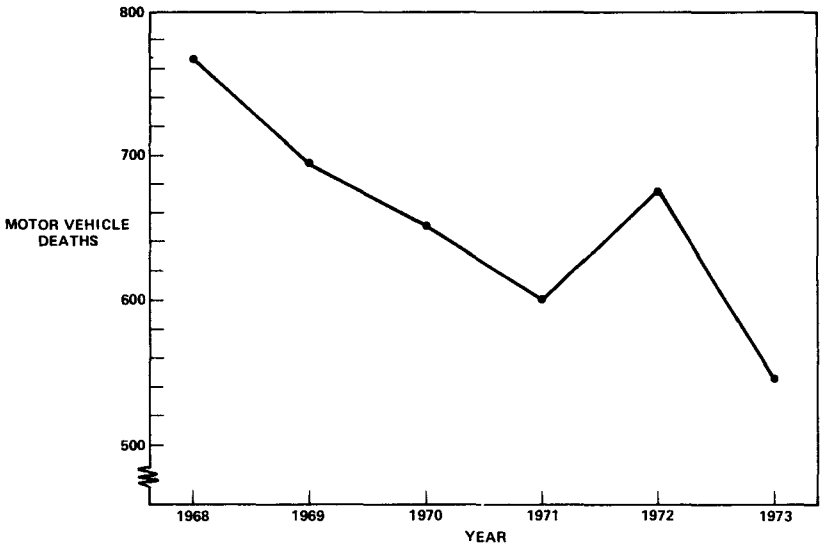


FIGURE 4. Thanksgiving holiday motor vehicle deaths*, 1968-1973.

Source: Insurance Institute for Highway Safety, 1973.

*IMMEDIATE DEATHS, THOSE OCCURRING BY MIDNIGHT ON THE LAST DAY OF THE HOLIDAY PERIOD.

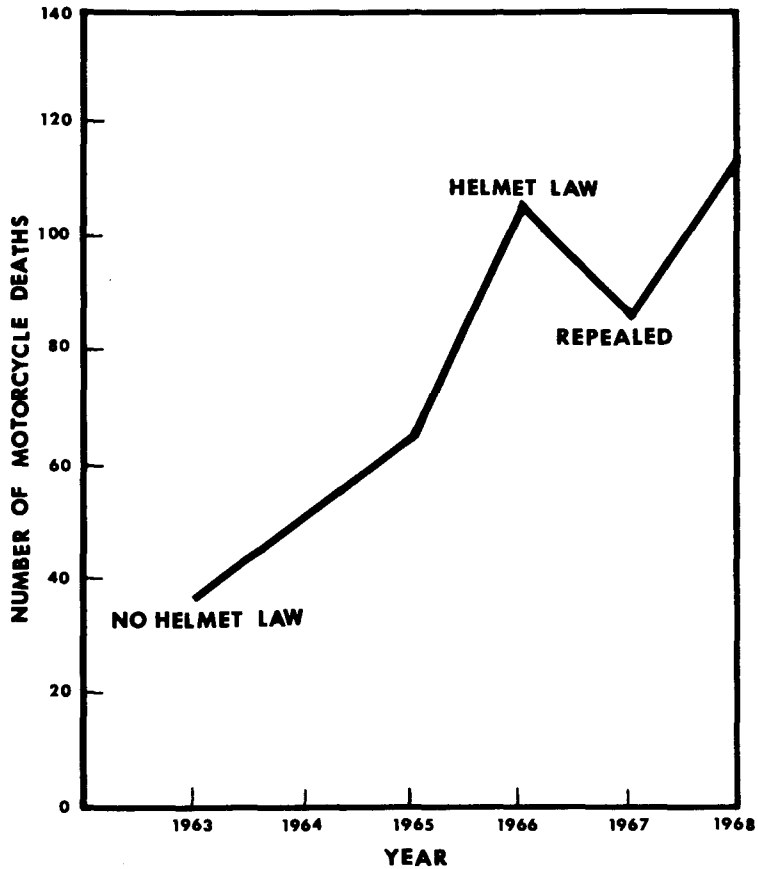


FIGURE 5. Michigan motorcycle deaths, 1963-1968.

Source: Klein & Waller, 1970.

is failed or refused a blood test may be demanded. If the blood test reveals an alcohol concentration of .08 percent or higher, the driver is deemed guilty of a crime, the punishment for which includes a mandatory license suspension for a year. The legislation is widely known both because it was a major political issue in its enactment, and because many claims of effectiveness were made for it shortly after its inception.

The initial claims for effectiveness of the British Road Safety Act were, as in the Connecticut case, based on before-and-after data, and were scientifically unacceptable. I went to Britain and brought home a variety of time series, nearly all of which were drawn from public records. In Figure 6, the interrupted time-series method is utilized to test the hypothesis that the legislation reduced traffic fatalities. (The data have been corrected to eliminate the effect of seasonal cycles and differential lengths of months.) Although the decline following the inception of the Act in October 1967 may appear small, it is actually rather impressive considering the fact that traffic deaths have many causes, and it is highly significant statistically. Note, furthermore, the absence of any peak in the curve just prior to the Act, eliminating any concern we might have about regression. Unlike the Connecticut case, application of interrupted time-series analysis to the matter confirms claims that the legal change caused a decline in fatalities.

Our confidence in and understanding of the cause-and-effect hypothesis can be increased in this case by contrasting data for times when, if the hypothesis is correct, the effect would be greater, and for times when it would most likely be smaller. This comparison introduces an expanded research model, the multiple time series, consisting of two or more simultaneous interrupted time series. It permits a control for history, the one factor previously mentioned which is not controlled in the simpler model. If the effect appears where it is expected according to the research hypothesis, and not where it is not expected, our confidence in the hypothesis is increased because a competing historical explanation would have to produce the same expectations and such a coincidence seems unlikely. The comparison is presented here in Figures 7 and 8, and it is most convincing. Figure 7 shows fatalities and serious injuries, appropriately corrected, for weekday

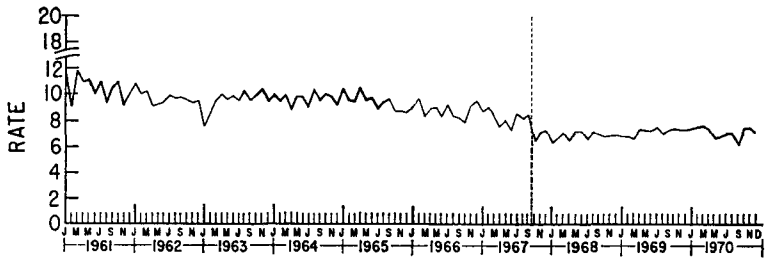


FIGURE 6. British fatality rate, corrected, seasonal variations removed, 1961-1970.

Source: Ross, 1973.

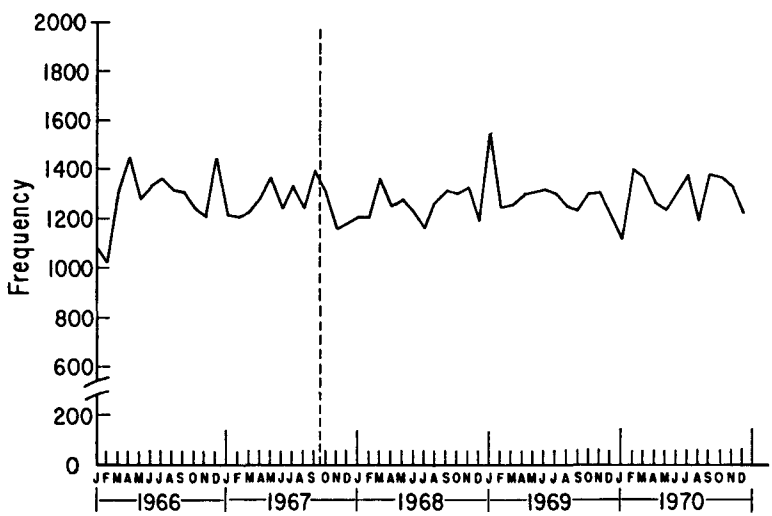


FIGURE 7. Fatalities and serious injuries combined, weekday commuting hours in Britain, 1966-1970.

Source: Ross, 1973.

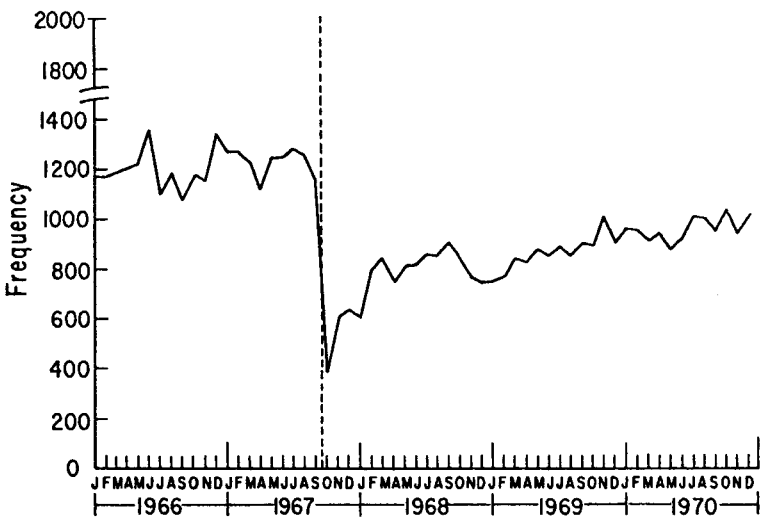


FIGURE 8. Fatalities and serious injuries combined, weekend nights in Britain, 1966-1970.

Source: Ross, 1973.

commuting hours when pubs are closed and social customs do not encourage drinking. There is virtually no change. Figure 8 presents similar data for weekend nights. The change is enormous, and is completely in conformity with the idea that it was produced by a law affecting drinking and driving. These series prove beyond reasonable doubt that the British Road Safety Act of 1967 was impressively effective in reducing highway casualties, although they do not in themselves explain how the reduction came about.

Another example of evaluation of drinking and driving laws achieved by interrupted time-series analysis is presented in Figure 9, which records the experience of eight of the American Alcohol Safety Action Programs (ASAPs) with two years of operation by 1972 (United States Department of Transportation, 1972). The figure is visually impressive, and the summary report from which the figure is drawn states that the change is statistically significant. Interpretation of this result is rendered rather difficult, however, by the great diversity of the ASAPs, each of which is a complex program involving various legislative, enforcement, judicial, and penal innovations. Moreover, the more numerous newer ASAPs when analyzed by the same method, in Figure 10, show no evidence of effect for the programs. Perhaps something effective is being done by some of the older programs but not the average newer program. The ASAP undertaking was not well designed for overall evaluation, and these two figures in juxtaposition do not permit any general conclusions at this time.

Several smaller scale attempts have been made to evaluate specific laws concerning drinking and driving. Unfortunately, in the smaller jurisdictions with lesser population bases the time series of fatal accidents tend to be very unstable. In this situation, the effect of legislation would have to be considerable for the interrupted time-series analysis to dispose of instability as a rival hypothesis, and in none of those cited here could the conclusion of effectiveness be supported. Figure 11 is typical; it is drawn from a study of Judge Raymond K. Berg's Chicago crackdown in December 1970 (Robertson, Rich, & Ross, 1973) promising mandatory seven-day jail sentences to drinking drivers. This study also produced a comparison series composed of fatalities in Milwaukee, which had no policy changes but where the drop at the time of the

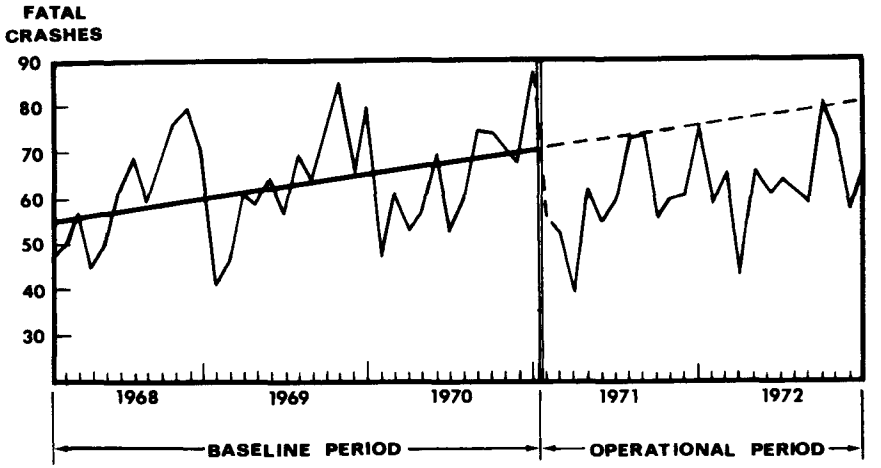


FIGURE 9. Total fatal crashes for 8 ASAPs with two full years of operation.

Source: U. S. Department of Transportation, 1972.

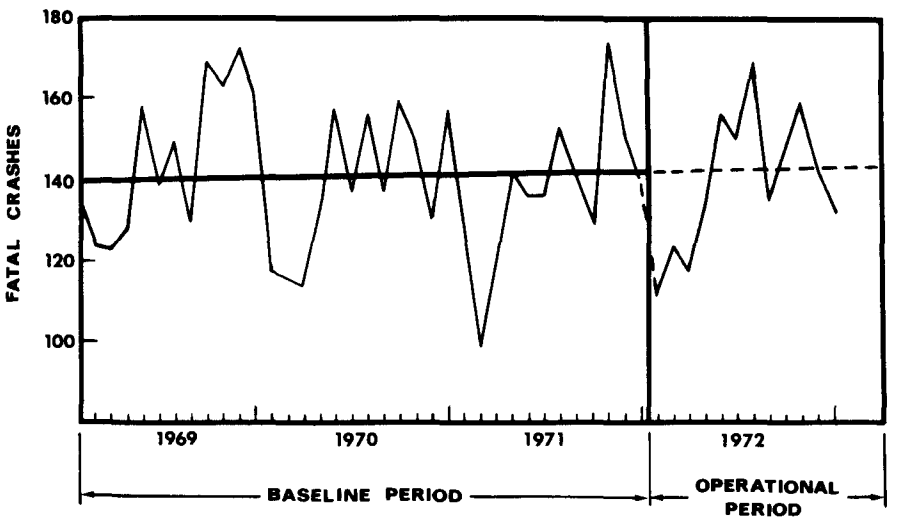


FIGURE 10. Total fatal crashes for 19 of 21 ASAPs with one full year of operation.

Source: U. S. Department of Transportation, 1972.

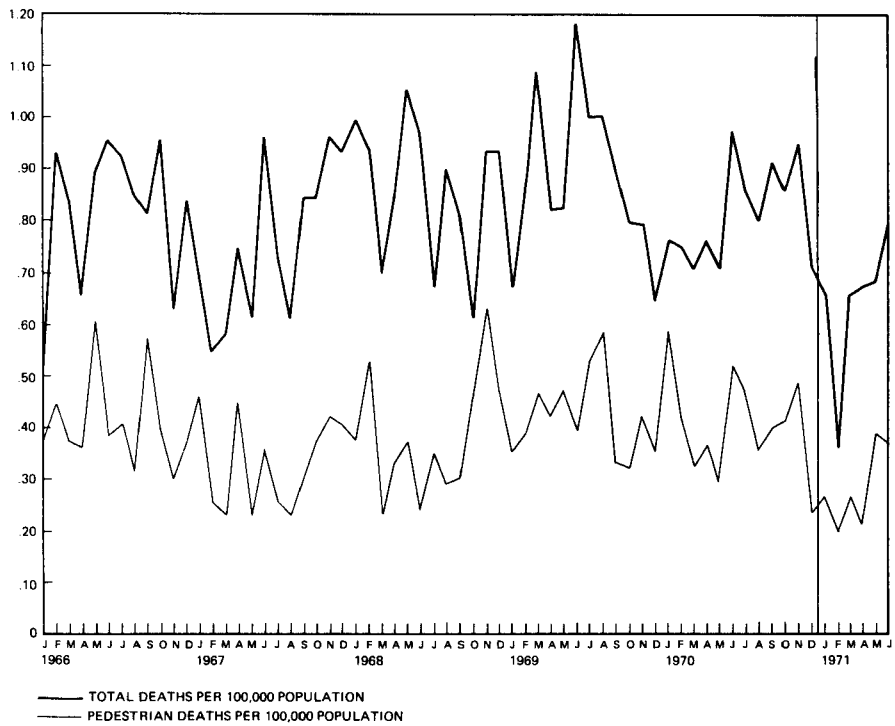


FIGURE 11. Chicago fatality rate, 1966-mid-1971.

Source: Robertson, Rich, & Ross, 1974.

Chicago crackdown was even greater than in Chicago. This negative evidence from the multiple time series, along with the observation that most of the drop in the Chicago data appeared to be in pedestrian fatalities (pedestrians were not the object of the legal threat), led the researchers to conclude that Judge Berg's claims of effectiveness were unsupported by the evidence. Similar results were obtained by Hunvald & Zimring (1968) in a study of the implied consent law in Missouri, and by Shover, Bankston, & Gurley (1974) in a study of drinking and driving legislation in Tennessee.

In view of the lack of consistent positive results in studies of laws related to drinking and driving, it is worthwhile considering the British evidence in more detail. Taking for granted the fact of effectiveness, four possible mechanisms seem plausible, and two of these can be investigated by means of interrupted time-series analysis. First, it is possible that British drivers drove less after the Act although combining drinking and driving in the same proportion as before. Second, the British might have consumed less alcohol, though not necessarily less in proportion while driving. Third, though drinking and driving as before, the drivers might have been more careful in order to avoid giving occasion to be tested, thus incidentally reducing accidents. Finally--certainly the intent of the law--both drinking and driving might have continued as before, but separated in time and place.

The first of these explanations was actually accounted for in Figure 6, which presented a rate based upon mileage and which therefore made allowances for the possibility of decreased travel. However, direct time-series evidence of estimated mileage can be presented, as in Figure 12. It is clear that no drop accompanied the legislation in October of 1967. (The yearly cycles typical of traffic data of all kinds show clearly in this figure, which has not been corrected to eliminate them.)

The second explanation is quite plausible, and was strongly feared by brewing and liquor interests in Britain during the national debate on the legislation. It is demonstrated to be false by the time series of releases from bond of spirits and beer, in Figure 13. No change is associated with the in-

ception of the Road Safety Act, although anticipation of a tax increase on spirits is clearly indicated in the spring of 1968.

No time-series data are available which bear directly on the third explanation, but for several reasons we can conclude that it is less plausible than the fourth, officially desired, mechanism. Survey data are available showing that in January 1968 considerably fewer drivers reported ever drinking and driving than in September 1967, and of those who drank outside the home more reported walking to the drinking place. Also in accord with this explanation is the fact that the proportion of traffic fatalities tested for blood alcohol with concentrations over .08 percent decreased from 25 percent in December 1966 through September 1967 to 15 percent in the corresponding period a year later. One can further speculate that drivers with blood alcohol concentrations in excess of .08 percent are unlikely to be able to control their behavior, in accord with the third explanation, so as to avoid accidents. We may thus conclude that the most likely explanation for the effectiveness of the Road Safety Act of 1967 is that British drivers were deterred from driving after drinking, but not from either drinking or driving as independent activities.

Unfortunately from the policy viewpoint, the time series we have cited also provide some evidence for the proposition that the deterrent effect of the British legislation was not permanent. The evidence is visible in the post-inception slopes of the curves in Figures 6 and 8 which, if extrapolated, suggest the likelihood of an eventual return of fatalities and serious injuries to previous levels. I have argued in the more detailed report on this legislation (Ross, 1973) that the diminished effectiveness of the law over time was due to its relatively low level of enforcement, which was in turn a consequence of public relations fears on the part of the police. Deterrence, I argue, depends strongly on a high perceived probability of apprehension. In the British case this impression was fostered in part by the enormous publicity attending the enactment of the legislation, but it was not subsequently supported in the behavior of the police.

Interrupted time-series analysis also helps us to investigate some unanticipated and unintended effects of legal reforms. Most of the policies I have mentioned have had as an

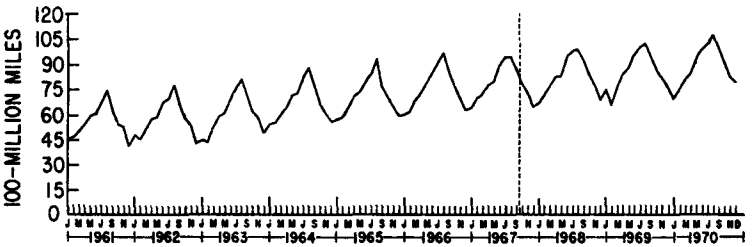


FIGURE 12. Estimated vehicle mileage in Great Britain, 1961-1970.

Source: Ross, 1973.

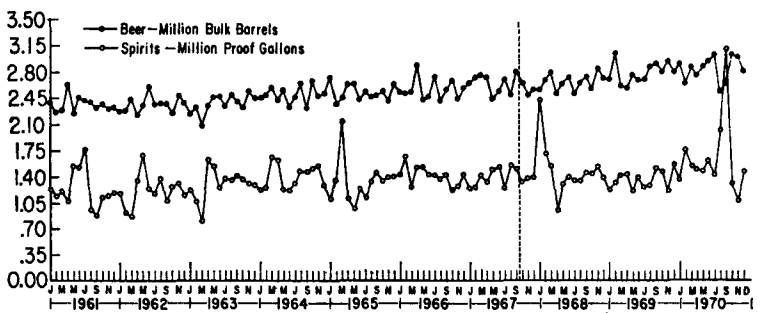


FIGURE 13. Releases from bond of beer and spirits in Great Britain, seasonal variations removed, 1961-1970.

Source: Ross, 1973.

important aspect an increase of severity of punishment, as compared with traditional penalties for the behavior in question. A predictable consequence of increasing penalties in an integrated legal system is a reaction on the part of that system to minimize the disruption of traditional relationships. The following figures illustrate changes in various of the examples studied. Figure 14, for instance, drawn from the Connecticut speed crackdown study, shows an impressive decline in arrests for speeding following the legal change. Although Connecticut officials interpreted this entirely as a result of more law-abiding behavior on the part of motorists, it is quite likely that it reflects in part a tendency of the police to charge marginal speeders in other legal categories, such as careless driving, or to issue no citations in the marginal case. Figure 15 shows that Connecticut judges became more reluctant to find accused speeders guilty, even though the charged violations were very likely more clear cut because of police discretion. The same phenomenon is illustrated in Figure 16, from the Chicago study, which shows that where judicial discretion was most available--where no chemical test for blood alcohol had been taken--there was a decline in convictions. The Phoenix ASAP provides the abbreviated time series of Figure 17, showing the same phenomenon with several different measures following an increase in severity of penalty for DWI (United States Department of Transportation, n.d.). Finally, Figure 18 from the Connecticut study indicates that resistance to the penalties increased among the penalized, as larger numbers of suspended drivers appeared willing to assume the risks of driving without a license. To be fair, it should be noted that few of these phenomena could be seen in the British data, but I believe the reason is that they were compensated by the fact that the chemical testing provisions of the Road Safety Act effectively controlled much of the discretion present in the legal system under British conditions.

CONCLUSION

The interrupted time-series method seems well suited for evaluating legal changes. As a quasi-experimental technique, it does not require random assignment of individuals or groups to different legal treatments, and it avoids the related ethical and practical problems which frequently interfere with at-

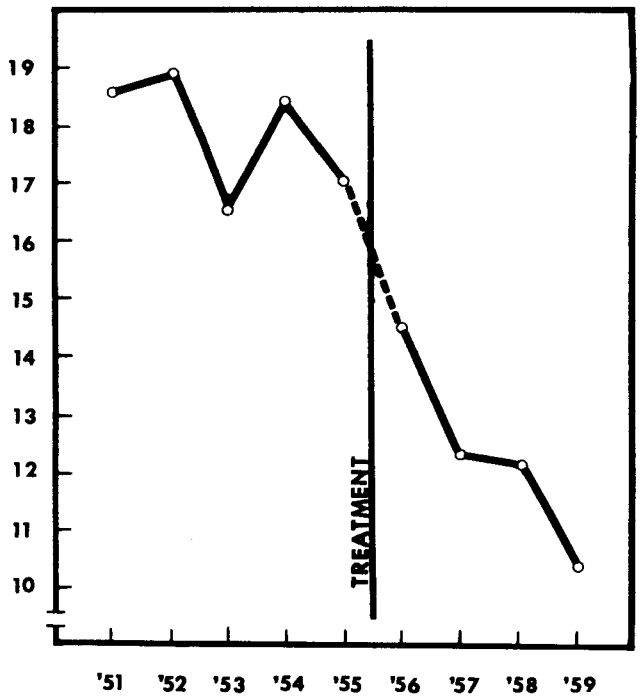


FIGURE 14. Speeding violations in Connecticut as a percent of all traffic violations, 1951-1959.

Source: Campbell & Ross, 1968.

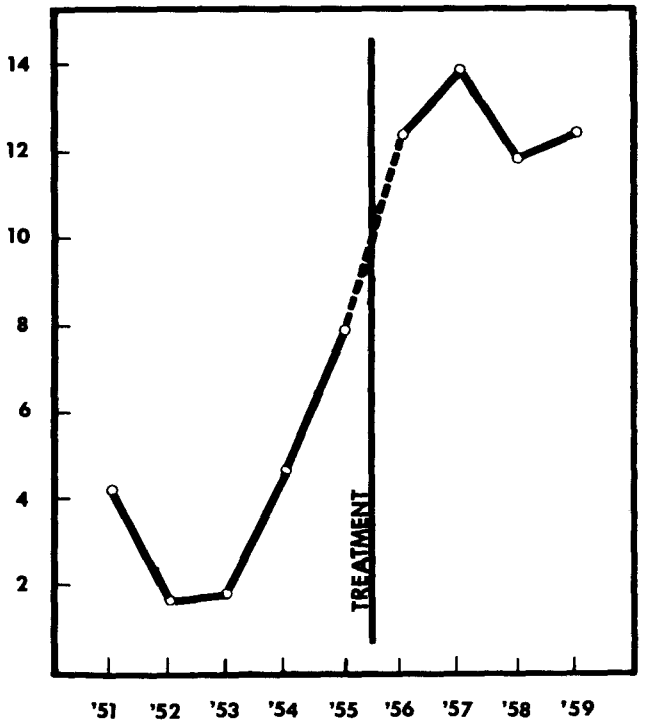


FIGURE 15. Percent of speeding violations judged not guilty in Connecticut, 1951-1959.

Source: Campbell & Ross, 1968.

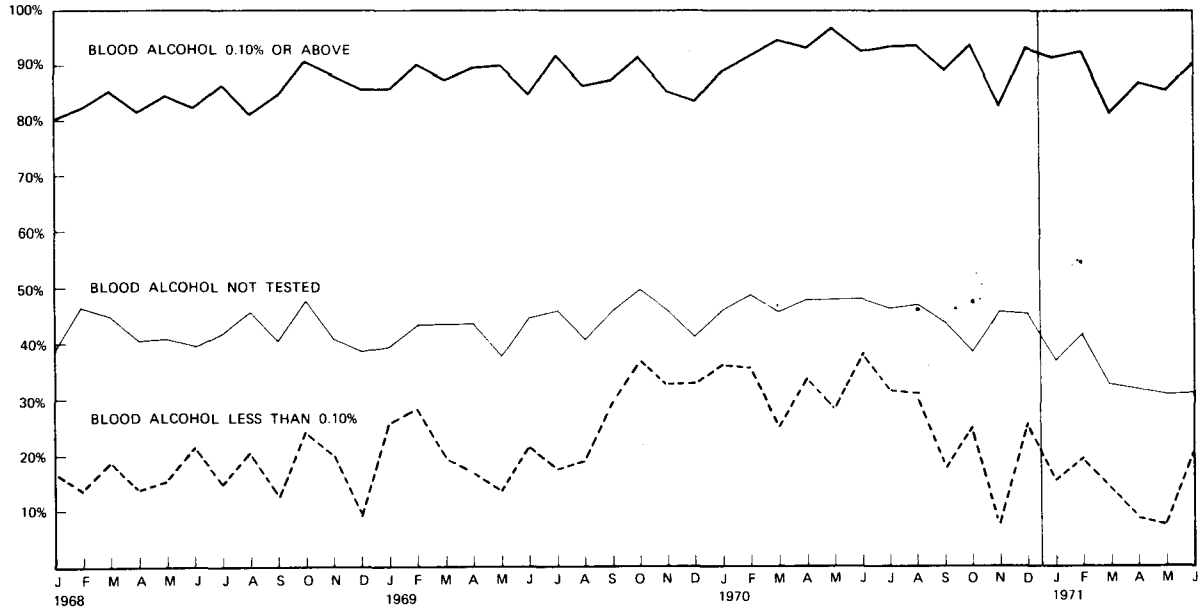


FIGURE 16. Percent of arrests resulting in conviction for driving while intoxicated in Chicago, 1968-1971, by blood alcohol concentration.

Source: Robertson, Rich, & Ross, 1974.

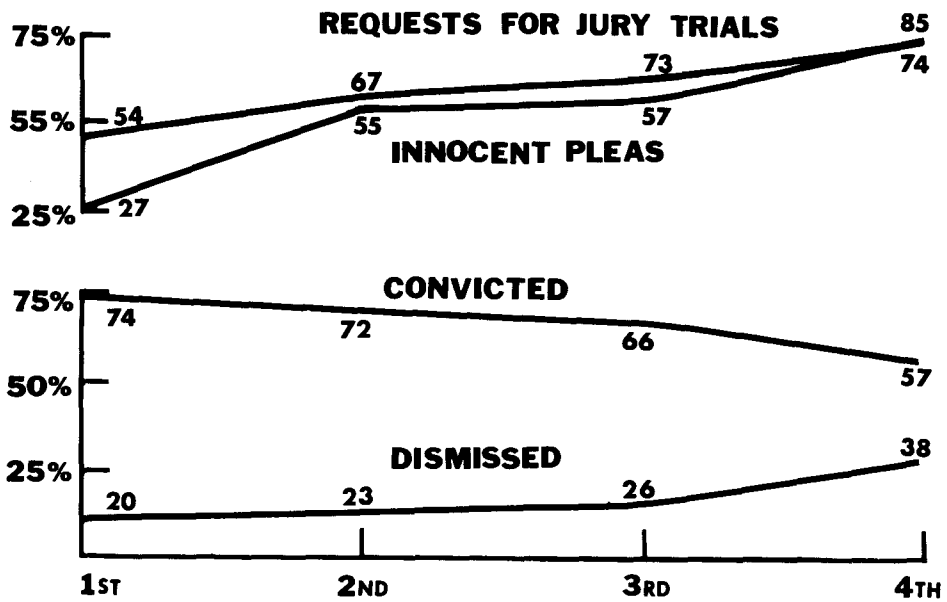


FIGURE 17. Requests for jury trials, innocent pleas, convicted and dismissed, reported by Phoenix ASAP, 1972.

Source: U. S. Department of Transportation, n.d.

We attempted to locate the original source of this document, but were not successful. However, it is our assumption that the interruption in time series occurred at the point of the third quarter in the calendar year. --- Editor's Note

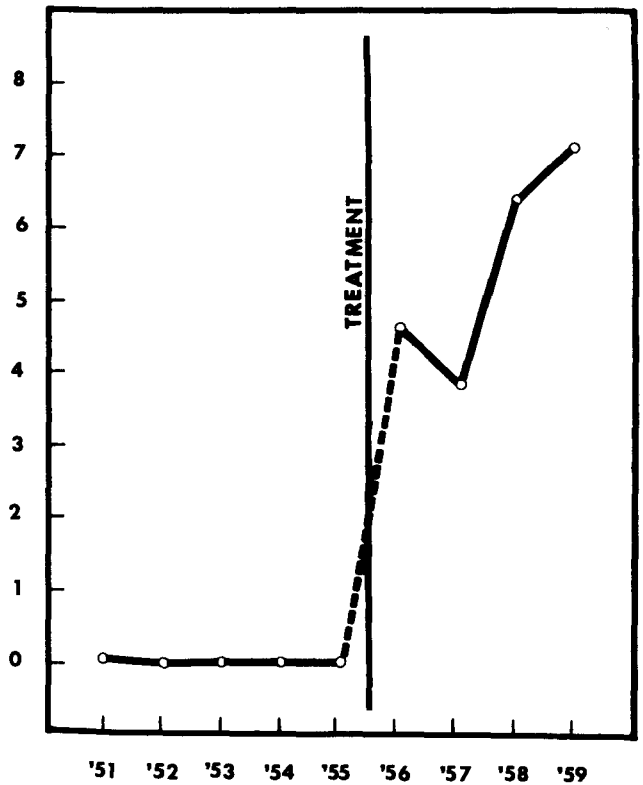


FIGURE 18. Arrests for driving with a suspended license, as percent of all suspensions in Connecticut, 1951-1959.

Source: Campbell & Ross, 1968.

tempts at experimentation in the law. In comparison to the inferior before-and-after study, which is most frequently used in practice, the interrupted time series is far more interpretable, and its data requirements are often met with easily available public data. All that is necessary is a routine series of measures of the effect variable over time. Where the administrator, and hence the investigator, is clear concerning the goals of the legislative program, appropriate measures may well be found. In the area of traffic law, fatal accidents are well measured, and we have long series for a variety of jurisdictions. Other types of accident are more frequent and hence provide larger data bases, but may have considerably less validity. In areas of law other than traffic, routine series of statistics are also commonly available, although their validity may be problematic in some cases. Examples are crime statistics, divorces, prices, population data, etc. Although to date most legal studies using interrupted time series have been in traffic law, there are some excellent studies in other legal areas, including divorce (Glass, Tiao, & Maguire, 1971) and welfare (Baldus, 1973).

Interrupted time-series analysis is, of course, no panacea for the needs of evaluation research. However, some suggestions can be offered that will increase the chances for successfully applying the model to the study of projected reforms. I would like to close with the following suggestions, based on those formulated by Donald Campbell (1969).

1. The projected reform should be introduced suddenly and totally. The more gradual the introduction, the more difficult it is to distinguish the effect of the reform from other possible sources of change in the effect variable. This condition is usually well met in studies of legal policies, which tend to have sharply defined inceptions.

2. The method depends on routine measurement, and any reform must not extend, at least initially, to the record-keeping system. One of the principle problems in interpreting the ASAPs has been that many of them included record-keeping reforms in their designs. If more drinking drivers were observed and reported under the reformed systems, then actual reductions in drinking and driving due to other activities would be hidden by the apparent increase due to what

we term instrumentation.

3. A program to be evaluated by interrupted time-series analysis must not be introduced in times of crisis, for fear that any effect would be confused with regression, which we would expect in the presence of extreme levels of a problem. The best way of accommodating to this necessity is to delay the reform until the level of the problem becomes more typical.

4. The analysis is more convincing where it utilizes multiple time series rather than just a simple interrupted time series. One potential source of comparison series is external--adjacent and similar jurisdictions. Although not discussed here, comparison series for use in the Connecticut study were created from the rates of adjacent and similar states, and the ASAPs were frequently compared with the balance of the states in which they operated (a bad choice when the ASAP was in the major metropolitan area of an otherwise rural state like Colorado). As was told, the Chicago data were more easily interpreted by the use of Milwaukee data. Another source of comparison series is internal, as demonstrated here by the comparison of weekday and weekend figures in Britain. Perhaps one of the latent benefits of American federalism is the ease with which comparative data can be introduced into our social research. In any event, the researcher should endeavor to broaden his research focus to include such data if the comparison seems at all reasonable.

In this paper I have tried to show the usefulness of interrupted time-series analysis in a variety of research contexts. I wish to emphasize that it seems especially suitable for the study of legal policies, where its requirements are often easy to satisfy and where adequate research models are often inapplicable. Its applicability to the study of traffic law, and particularly the study of laws relating to drinking and driving, is enhanced by the relatively clear goals for legislation in this area. I hope and predict that the interrupted time series will become an increasingly useful and utilized research method.

ACKNOWLEDGMENTS

The helpful comments of Gene V. Glass and Murray Blumenthal are gratefully acknowledged.

REFERENCES

Baldus, D.C. Welfare as a loan: An empirical study of the recovery of public assistance payments in the United States. Stanford Law Review, 1973, 25, 125-250.

Campbell, B.J. Highway safety program evaluation and research. Traffic Digest and Review, 1970, 18, 6-11.

Campbell, D.T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D.T., & Ross, H.L. The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. Law and Society Review, 1968, 3, 33-53.

Campbell, D.T., & Stanley, J.C. Experimental and quasi-experimental designs for research in teaching. In N.L. Gage (ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Glass, G.V. Analysis of data on the Connecticut speeding crackdown as a time-series quasi-experiment. Law and Society Review, 1968, 3, 55-76.

Glass, G.V., Willson, V.L., & Gottman, J.M. Design and Analysis of Time-Series Experiments. Boulder: Laboratory of Educational Research, University of Colorado, 1972.

Glass, G.V., Tiao, G.C., & Maguire, T.O. The 1900 revision of German divorce laws: Analysis of data as a time-series quasi-experiment. Law and Society Review, 1971, 6, 539-562.

Hunvald, E.H., Jr., & Zimring, F.E. What happened to implied consent? Missouri Law Review, 1968, 33, 325-399.

Klein, D., & Waller, J.A. Causation, Culpability and Deterrence in Highway Crashes. Washington, D.C.: United States Government Printing Office, 1970.

Robertson, L.S., Rich, R.F., & Ross, H.L. Jail sentences for driving while intoxicated in Chicago: A judicial action that failed. Law and Society Review, 1973, in press.

Ross, H.L. Law, science and accidents: The British Road Safety Act of 1967. Journal of Legal Studies, 1973, 2, 1-78.

Ross, H.L., & Blumenthal, M. Some problems in experimenting in a legal setting. Unpublished multilithed manuscript, available from the authors, 1974.

Shover, N., Bankston, W.B., & Gurley, J.W. Increasing sanctions, citizen awareness, and traffic death: The experience in another jurisdiction. Unpublished multilithed manuscript, available from Department of Sociology, University of Tennessee, 1974.

Sween, J., & Campbell, D.T. A study of the effect of proximally autocorrelated error on tests of significance for the interrupted time-series quasi-experimental design. Unpublished multilithed manuscript, available from the author, 1965.

United States Department of Transportation, National Highway Traffic Safety Administration, Office of Alcohol Countermeasures. Alcohol Safety Action Projects: Evaluation of operations--1972. Washington, D.C.: United States Government Printing Office, 1972.

United States Department of Transportation, National Highway Traffic Safety Administration, Office of Alcohol Countermeasures. A review of the case for and against mandatory jail sentences for a first conviction of driving while under the influence of alcohol. Unpublished mimeographed manuscript, available from the agency, n.d.

Zylman, R. Drivers' records: Are they a valid measure of driving behavior? Accident Analysis and Prevention, 1972, 4, 333-349.