University of North Carolina Highway Safety Research Center

e-archives

Lindsay I. Griffin, Brain Powers, and Catherine Mullen. (1975). Impediments to the Evaluation of Highway Safety Programs. Chapel Hill, NC: University of North Carolina Highway Safety Research Center.

ATTENTION

The enclosed report is a reprint of the original technical report which has recently gone out of print. Its content does not differ in any way from the original report. The format differs slightly due to time restrictions in the reprinting process.

We hope that this report will fulfill your interests. We appreciate your continued concern in highway safety.

ABSTRACT

In the U.S., the death rate per 100 million vehicle miles has dropped three-fold in 40 years -- from 15.60 in 1933 to 4.30 in 1973. Although this decline has been attributed to numerous and varied regulations, programs, and countermeasures, the real reasons for the decrease are unknown. In order that realistic decisions regarding the continuation, addition, or deletion of highway safety programs be made in the future, it is imperative that valid evaluations be conducted. Without such evaluations, the effectiveness of highway safety programs cannot be determined. If this determination is not made, limited available funds cannot be allocated to those programs which are most effective in saving lives and reducing injuries and property damage.

Although the evaluation process has been criticized on numerous grounds, real "effectiveness" evaluations have rarely been carried out. The most frequent and crucial impediments in the area of highway safety evaluation are seen to be: (1) a lack of understanding of evaluation, (2) an unwillingness to have programs undergo evaluation if an understanding of the process does exist, (3) a paucity of trained personnel to carry out evaluations, and (4) the existence of inadequate tools, procedures, and data bases for establishing sound evaluative research procedures.

Recommendations are made detailing ways in which these deficiencies might be overcome and future highway safety evaluations improved.

This study was prepared for the Motor Vehicle Manufacturers Association of the United States, Incorporated, and was supported with funds under MVMA Agreement Number UNC 7404-C5.1

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF FIGURES

ACKNOWLEDGMENTS

Special appreciation is expressed to the National Association of Governor's Highway Safety Representatives and to the many Governor's Highway Safety Representatives who were interviewed by the authors. Noel Bufe, formerly Executive Director of Highway Safety Planning, Michigan State Police was particularly helpful in contributing ideas and information.

The assistance of the following research personnel is also gratefully acknowledged: Ronald Coppin, Chief of Research and Statistics of the California Department of Motor Vehicles, with a special note of thanks to Raymond Peck and Richard Harrano of the Research and Statistics Branch; Noel Kaestner, Consulting Psychologist with the Oregon Traffic Safety Commission; Charles J. Keese, Director of the Texas Transportation Institute at Texas A & M University; Richard Williams, Project Director, and Barbara Moser, Research Associate, with the Institute for Human Ecology in Raleigh, North Carolina; and Patricia Rieker, Assistant Professor, U.N.C. Department of Sociology.

Among the staff at the U.N.C. Highway Safety Research Center, appreciation is expressed to Forrest Council, Elizabeth House, William Hunter, John Lacey, Donald Reinfurt, and Patricia Waller for their contributions to this report. The authors owe special gratitude and recognition to B.J. Campbell. His former and present work in the field of highway safety in general, and highway safety evaluation in particular, have proved to be an invaluable base upon which to build the content of this report.

Finally, sincere appreciation is due to the Motor Vehicle Manufacturers Association for providing financial support for this project and to Mr. David Willis, Economic Research Department of MVMA, for his continual assistance to the authors throughout the duration of the project period.

I. INTRODUCTION

Historical Context

The U.S. now has the lowest death rate per 100 million vehicle miles of any industrialized nation in the world. Although the death rate has dropped three-fold in 40 years, the reasons for the decline are not understood (Accident Facts, 1974). Educated guesses can be made, but, because of lack of records and the vast number of programs claiming some credit for the decrease, the real reasons will probably never be determined. A multitude of changes have affected the accident situation in the last 40 years. Because the situation has been dynamic with many and diverse changes taking place, there has been very little time or interest in measuring these efforts. The hue and cry has been "Do something about the carnage on the highways," not "Find out if what is being done makes any sense."

"The components generally listed as comprising the essential segments of what is called 'highway use' are (1) the highway, (2) the vehicle, (3) the driver, and it is commonly recognized that these components cannot be compartmentalized into neat categories, for there is much interplay between all of them" (Reese, 1965, p. 1). While the dramatic decrease in death rate is certainly primarily due to improvements in these areas, the decline may also be attributable, at least in part, to improved medical services over the years. Not only have there been improvements in emergency medical service (EMS), but also in hospitals, equipment, and personnel.

It should further be realized that the tremendous decline in death rate which has been recorded in the last four decades may not be as large as is presumed to be the case. The decline may be partially artifactual. Griffin (1974) acknowledges this possibility when he points out that as a society becomes more urbanized, the denominator of death rate, "100 million vehicle miles," changes character. In rural societies, "100 million vehicle miles" are accumulated as a relatively few vehicles travel long distances, presumably at relatively high speeds.

In urban societies, "100 million vehicle miles" are accumulated as a relatively large number of vehicles travel short distances, presumably at slower speeds. To the extent that population density influences the conditions under which vehicle miles are accrued, so, too, density influences death rate. As the United States has become more densely populated, relatively more vehicle miles have been logged within cities, at lower rates of speed. Accordingly, the motor vehicle death rate should have declined over the years -- even if the driver, the vehicle, and the roadways had remained unchanged.

Today, the death rate is lower than in previous years. Why? No one really knows. Many countermeasures claim credit for the reduction, but the relationship between the various countermeasures and the declining death rate remains speculative, unproved, and dubious. Because past safety programs were not accurately and systematically evaluated, "common sense" assessments of highway safety programs have prevailed. In lieu of facts, a rich mythology has arisen concerning the effectiveness of certain countermeasures. Greenshields (1970) argues that "a major deterrent (to highway progress) is an established way of thinking. There are persistent myths that seemingly cannot be exorcised" (p. 674). Suchman (1967) states that "the arbitrary selection of problems and services tends to stress traditional activities at the expense of newly developing areas" (p. 16). And Jacobs (1961) insists that "In the absence of an experimental attitude towards accident therapy...it would appear that any success in this field will be more the result of good fortune or happy speculation than knowledge and understanding" (p. 21).

For the last 40 years, the field of highway safety has suffered from the absence of program evaluation. Today the situation is somewhat improved, but much remains to be done.

What is Evaluation?

The basic function of evaluation is to make judgments of worth. These judgments result in studies ranging from those utilizing the most rigorous, quantitative, experimental designs to those involving the most capricious and subjective of estimates. "And with respect to the growing confusion about what evaluation is, Columbia's Carol Weiss has termed it well -- as a rubber word that is stretched to mean whatever you want it to mean" (Davis, 1972, p. 3-4). Granted that evaluation is a very elastic word and that it also has a generally positive connotation (in the sense that it is seen as "scientific"), it is possible that many procedures will attempt to masquerade as evaluations. "If you can't approximate an evaluation, have no fear," standard practice tells us, "Just call whatever you have done an evaluation regardless."

Since evaluation is an ambiguous concept for most people, it is very important to explain exactly what the term means. Essentially, there are three types of evaluation:

Type I (Subjective, clinical assessment). This type of evaluation is reflected in the impressions of so-called "experts." It is usually devoid of numerical data and is necessarily filled with subjective feelings and opinions. In the field of highway safety, judges, police officers, and medical doctors frequently serve as expert evaluators.

Type II (Process evaluation). This type of evaluation is numerical in nature but devoid of dependent variables to monitor. It is used to determine if (and to what degree) the treatments or manipulations advocated at the initiation of a program were carried out.

Type III (Outcome, effectiveness evaluation). This type of evaluation is characterized by the measurement of a dependent variable. When a program is instituted which is intended to influence one or more other variables, then measurement of changes in those other variables constitutes a Type III evaluation.

To take a specific example, suppose that a remedial driving course has just been initiated to improve the driving skills and subsequent accident experience of a group of poor drivers. A Type I evaluation of the course would be performed by asking several educators to look over the course outline and to observe and comment on the degree to which the course should benefit the students. Alternatively, the students themselves might be asked to express their opinions of the course. A Type II evaluation of the course might consist of counting the number of students attending the course, estimating the cost of the course per student trained, etc. A Type III evaluation would compare the accident records of course graduates to the accident records of a comparable control group.

While the credibility of the first type of evaluation can be readily questioned, both Type II and Type III evaluations can be conducted appropriately under different circumstances. For example, one federal program standard directs each state to upgrade and improve its traffic records system. For programs initiated under this standard, Type III evaluations are inappropriate. Traffic records systems are not designed to have a direct influence on accident rates, death rates, property damage, driver knowledge, or any other conceivable dependent variable. Instead, a traffic records system should be looked upon as a

support system which in the long run may improve the highway safety picture, but which in the short run will never save a life or lessen the probability of injury. Traffic records systems should be evaluated with the Type II procedure. On the other hand, many programs which are initiated in the highway safety field are specifically designed with goals which can be measured. Alcohol-related programs, for example, are designed, "to achieve a reduction in those traffic accidents arising in whole or in part from persons driving under the influence of alcohol" (Highway Safety Program Standards).

For many programs, only Type II evaluations are possible, but for many other programs which are amenable to Type III evaluations, the latter type of evaluation should be preferred. Type II evaluations cannot serve in lieu of Type III evaluations; they can only supplement and add to the information which the effectiveness evaluation yields.

In the rest of this report, the term "evaluation" will refer to effectiveness evaluation unless specifically stated to the contrary.

Reasons for Evaluation

At this point it might be asked -- why should we evaluate our programs? After all, in the last 40 years our fatality rate has declined substantially. We must be doing something right!

Indeed, we must be doing something right. But without adequate evaluations, we do not know what those things are. We do not know which programs have been responsible for our successes in the past and, therefore, we do not know which programs should be implemented in the future.

Program evaluation should accomplish three major purposes: (1) It should determine whether or not the program is accomplishing the goals it was designed to accomplish. (2) It should determine how efficiently the program is accomplishing its stated goals. (3) It should determine if the program is producing results contrary to its goals.

Many highway safety programs have been initiated over the years which have not reduced the loss of life, limb, or property. Many such programs are still in existence, and still more will come into existence. With good evaluations, these programs which produce no benefit can be discovered, and hopefully eliminated.

By eliminating ineffective programs, more than the taxpayer's dollar is saved; lives are saved. The funds allotted by society to highway safety are pitifully finite. Funds spent on one project

represent funds unavailable for other projects. If Project A is ineffective, not only does it squander funds, but it prevents those same funds from being applied to Projects B, C, and D where some safety gain might be realized. By pointing out the ineffectiveness of Project A, evaluation can serve in a real (albeit indirect) way to save lives.

While few taxpayers would argue that ineffective programs should be maintained, many would argue that any program which saves a human life is beneficial. This philosophy that says "any life lost is a life worth saving" is a pernicious one; pernicious because the funds applied to saving one life, and only one life, might very well be applied in another way so as to save 10 lives. Funds should be allocated to highway safety countermeasures which will save the most lives, reduce the most injuries, and so forth. Only when the payoff function is known for the various countermeasures can safety dollars be rationally distributed. The same logic which says that no yield projects are costly in terms of life and limb dictates that low yield projects will also be costly (B.J. Campbell, 1970).

Some countermeasures which are initiated in the name of safety may not only be ineffective, they may actually be injurious. In a recent study of guardrails in Texas, it was concluded that many embankments that required guardrails under existing legislation did not actually need protection. It was found that, at certain speeds and for certain encroachment angles, it would be safer for the driver to run off the road than to hit existing guardrails. "Guardrails should be used for conditions where the severity of an errant automobile redirected by the guardrail is less than the severity of an errant automobile transversing the unprotected embankment" (Ross & Post, 1972, p. vi).

Unfortunately, the Texas study is not an isolated example. Findings of a recent evaluation study of crosswalks in San Diego revealed: "Unjustified and poorly located marked crosswalks may cause an increased expense to the taxpayers for installation and maintenance costs which may not be justified in terms of improved public safety. Indeed, such crosswalks may tend to increase the hazard to pedestrians and motorists alike" (Herms, 1970, p. 31). In this case, due to lack of pedestrian caution, there were twice as many pedestrian accidents at marked crossings as there were at unmarked crossings (when usage was taken into consideration).

Even seemingly innocuous countermeasures may be counterproductive in the long run. For years, the public was exposed to an expensive media effort concerning alcohol abuse: "If you drink, don't drive." This effort was not only a waste of time and money, but the message itself was distorted and, ultimately, probably had a deleterious effect.

The public often misunderstood what constitutes problem drinking (as opposed to social drinking), and sympathy was created for the drunken (DUI) offender -- "there but for the grace of God, go I" (Swinehart & Grimm, 1972).

These examples illustrate that evaluation is not simply an activity designed to eliminate ineffective programs. Occasionally, programs are instituted which are directly responsible for killing people. By identifying those programs which produce unanticipated consequences, and by discovering those programs of little or no benefit, evaluations serve the cause of highway safety.

Impediments to Evaluation

Having defined effectiveness evaluation and listed several reasons why effectiveness evaluations should be carried out, it can now be stated that relatively few effectiveness evaluations have ever been performed on highway safety programs. Those which have been performed are often riddled with error and fallacy.

In 1966, when the federal government became actively involved in highway safety, a series of 16 program standards (presently 18) was promulgated. These standards carry the weight of law and set minimal requirements toward which the states are obliged to move. The standards pertain to such safety areas as driver licensing, driver education, and motorcycle safety. Each state is allotted funds on a federal/state matching basis to finance programs and projects designed to improve its safety record. Programs financed by means of these funds must be shown to fall under one of the 18 standards, and each program must be evaluated. It is primarily the responsibility of the Governor's Highway Safety Representative in each of the several states to establish the priority of highway safety programs within his or her jurisdiction, to fund the programs, and to see to it that each program carried out under his or her auspices is evaluated.

In January 1974, letters were sent to Governor's Highway Safety Representatives in each of the 50 states, Washington, D.C., and Puerto Rico. In the letters, each representative was asked:

> If you have carried out evaluations of highway safety programs within your state and have issued reports on those evaluations, would you send us a copy?

In response to the 52 letters sent, 24 answers were received. Of these 24, only 12 provided examples of evaluations which had been conducted.

The following excerpts are taken from nine of the respondents who did not provide copies of evaluations:

1. Our office has no evaluation reports to offer nor do we know of any other agency which has.

2. We have done very little in highway safety evaluation.

3. I am sorry to advise we are only now in the process of organizing a truly effective routine for monitoring and evaluation.

4. I am sorry to notify that, at present, we have carried (out) no evaluation(s) for effectiveness in any of our programs.

5. _____ has not done any true evaluations for the highway safety program to date that has any documentation.

6. We are in the process of developing procedures for quantitative project evaluations at this time. We are unable to forward anything to you at this time that would be meaningful.

7. Regrettably, I must respond in the same manner as did most states to a similar inquiry I made several years ago. Although, I believe the overall situation has considerably improved, I know of no evaluations of programs or countermeasures in _____ at either the state or local level that I could recommend to you as a "model" evaluation.

8. A major problem confronting evaluation in _____ is the lack of base line data. Past efforts to obtain the data have been only partially successful and as a result our evaluations are liberally laced with personal judgments.

9. The _____ Office of Highway Safety has not published project evaluations in the past. Due to minimal Federal and State funding the highway safety projects that we have completed in the past have been relatively small and of short time periods.

There are numerous reasons why evaluations are not conducted. Some programs have not been evaluated due to lack of funding; other programs have not had the requisite personnel to carry out the evaluations; still others have not had appropriate data with which to work. The specific reasons why given programs have not been evaluated are as numerous as the programs which have been left unevaluated.

If evaluation is ever to become commonplace, if programs are ever to be funded on the basis of demonstrated merit, the myriad impediments to the evaluation process must be subsumed under several basic headings and addressed as groups. If each failure to evaluate a program is seen as the resultant of some unique set of impediments, then the state of the art of highway safety evaluation cannot be improved. Only by grouping impediments can recommendations be formulated which will alleviate whole categories of impediments at one time.

It should further be recognized that there is no right or wrong way to categorize impediments to the evaluation process. There are categorizations which are practical and expedient and which allow for recommendations which could redress the present deplorable state of highway safety evaluation; there are categorizations which are of some theoretical interest, but which do not generate any practical recommendations. The following categorization is offered as a delineation of the problem of impediments to highway safety evaluation. It is hoped that this classification will yield practical, specific, and concrete recommendations whereby evaluations might be more forthcoming, and of better quality (See Figure 1).

External impediments.

Naive ignorance.

Before evaluations in highway safety or any other discipline become commonplace, there must be a greater awareness of what evaluation is , what its strengths are, and wherein its weaknesses lie. At the present time, the public is not familiar with the process of evaluation, nor are they familiar with what effects evaluation could have on their tax dollar. Furthermore, the representatives of the public, namely members of the U.S. Congress and members of the state legislatures within the several states, do not fully appreciate what evaluation is or how it might be used in the rational allocation of public funds. Within the administrative branches of government, both at the federal and state level, there are staffs of bureaucrats charged with the responsibility of carrying out programs devised by their agencies. Often these programs are initiated on the basis of "common sense,"

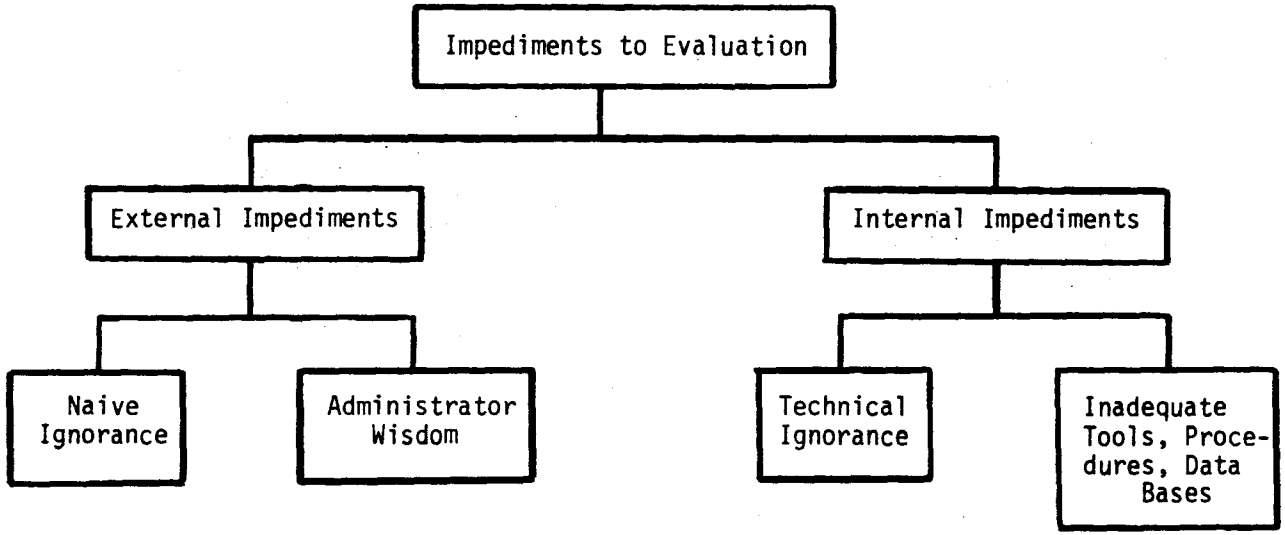


Figure 1. Impediments to evaluation.

"folklore," or the "prevailing view," without any thought being given to whether or not the program produces any benefit. The very fact that a program is initiated seems to imply that the program "works" and that "something is being done." This reasoning is obviously fallacious and, at the same time, wasteful.

Administrator wisdom.

Opposed to those individuals who have no conception (or very little conception) of what evaluation is, there exists another group which is acutely aware of the evaluation process and the consequences which it can have on ongoing programs. These individuals who are cognizant of how evaluation works realize: (1) that most highway safety programs are funded at a very low level when the goals of that program are taken into account, (2) that the dependent variables by which the effectiveness of a highway safety program is measured are very insensitive, and that therefore, (3) most highway safety evaluations have tended to produce negative results.

State administrators and program managers are not opposed to saving lives, reducing injuries, or protecting property, but they are opposed to program evaluations which they have good reason to believe will result in the verdict -- "no effect." The primary goal of administrators and managers is the sustenance and perpetuation of their organizations. Whether this goal is appropriate, indeed, whether this goal is morally defensible, is not an issue here. Practical experience tells us that administrators and managers, like the rest of us, tend to act in their own best interests.

From an administrator's point of view, evaluations are often seen as impediments to his interests and his mode of operation. Not only does a negative evaluation indicate that his program is of little worth, but such an outcome might possibly result in the curtailment of funds. To avoid such an unfavorable outcome, many administrators and program managers have resorted to prostitutions of the evaluation process itself. Rather than trying to determine in any accurate fashion whether or not a given program is effective, these individuals conduct "evaluations" which are little more than the subjective impressions of some program protagonist.

In short, if state administrators and program managers see little benefit in performing evaluations, and if the act of performing evaluations has a high probability of being detrimental, then it seems reasonable to expect that administrators will be less than enthusiastic about the evaluation process.

Internal impediments.

Technical ignorance.

Once a decision has been made to evaluate a program, once the goals of that program have been defined and the funds committed, there exists a real problem within the highway safety community of finding adequately trained technical people to carry out the evaluation. The number of erroneous highway safety evaluations which are now in print indicates the technical deficit in this field. The most basic errors of experimental design abound. Inadequate control groups, poor data collection procedures, non-randomized groups, are scattered throughout the literature. Effective programs are called worthless, and worthless programs are called effective.

Inadequate tools, procedures, and data bases.

If the decision has been made to evaluate a project and competent personnel have been enjoined to carry out the work, there still exists the real question of whether or not the evaluators will have sufficient funds, adequate tools, and the necessary authority to carry out valid evaluations. Evaluations are not performed in sterile laboratory settings, but within the scope of a project or program which is often very inhospitable to the requirements and demands of evaluation procedures. The obstacles which lie before the competent evaluator in the field of highway safety are many. The assurances of cooperation which he receives at the initiation of the project usually last only until the first hint of a negative finding. Data promised by a municipal police department, a Department of Motor Vehicles, or a hospital may not be forthcoming. Without adequate tools, procedures, and data, the most gifted evaluator will fall short of a valid evaluation.

II. EXTERNAL IMPEDIMENTS TO PROGRAM EVALUATION

A. Naive Ignorance

1. Many not aware of evaluation.

Evaluation is not a new concept. It has been in existence for several decades. It was derived from scientific method and is predicated upon the assumption that the effectiveness of a program can be determined only to the extent that the results (goals) of that program can be quantified and measured. Proponents of evaluation insist that program effectiveness cannot be defined by common sense, past practices, or political fiat. Empirical data are the only "facts" whereby the worth of a countermeasure can be ascertained.

Evaluation is not a new concept even within the field of highway safety. For the last twenty years, a small group of researchers has called for the evaluation of highway safety programs. For the last nine years the federal government has been committed to highway safety program evaluations. However, large segments of the public are even now totally ignorant of the evaluation process. Many legislators and administrators have very little idea of what evaluation is all about.

Rather than asking whether or not a certain program is effective, many citizens, legislators, and administrators feel that they know which programs are effective. They know that drunks should be "gotten off the road," and, consequently, they are in favor of breath test programs, license suspensions, or jail sentences for DUI offenders, etc. They know that "speed kills" and consequently they are in favor of increased police enforcement. They know that the program which they advocate works, and so the notion of evaluation never occurs to them.

If these people were constrained to carry out evaluations on their pet projects, they would not object. However, they would probably see the evaluation as a waste of time and money. If perchance the evaluation did not substantiate what they "knew," they would doubt the quality of the evaluation, not the project.

2. Expert opinion.

Those advocates of highway safety programs who are a little more conservative and a little less sure of the benefits to be derived from the program that they advocate sometimes call in "expert witnesses" to make judgments of worth on the program in question. The quality of the opinions yielded by these "expert witnesses" varies considerably. Individuals with little or no training in the field of highway safety and little or no training in evaluation methodology are asked to predict whether or not a given program saves lives. Often these judgments are qualitative in form and based on few, if any, references to other similar programs.

While it should be readily conceded that some individuals are quite expert at making clinical judgments concerning highway safety programs, it should also be recognized that most experts do little more than serve the cause of the status quo. It should further be recognized that it is difficult to discriminate between those experts who are capable, and those who render opinions which are without basis and often wrong.

The usual criteria for choosing highway safety experts are academic training and professional experience. It is somehow assumed that those individuals who have acquired medical degrees or professional degrees have simultaneously acquired the ability to judge the potential effectiveness of a program. Similarly, police officers who have served on a force for 20 or 30 years are also assumed to have attained this ability. While some medical doctors, judges, and retired police officers are sound of highway judgment, most are totally incompetent to render judgments in the area.

Expert opinion on a program constitutes the minimal level of program evaluation. In this report, this type of evaluation has been referred to as a Type I evaluation, and it has been stated that the validity of Type I evaluations is open to considerable question. Those advocates who resort to this type of evaluation usually do so with the best of intentions. They do so in the belief that the opinions of experts will substantiate the benefits to be derived from that program which they would like to see instrumented or continued. They choose this form of evaluation, not to avoid the more rigorous Type II and Type III evaluations, but in the mistaken belief that what they are doing is evaluation. Their error is not one of intent, but naiveté.

3. Evaluation, endorsement, and personal judgment.

At a slightly more advanced level than those who are totally ignorant of the evaluation process, there exists another group which is willing to entertain evaluations, but which is unwilling to accept the outcomes of evaluations which prove contrary to preconceived notions of effectiveness. These people use evaluation when it serves their purposes and substantiates their claims. They reject it as inappropriate or inadequate when a negative finding emerges.

The administrator who is looking for endorsement rather than evaluation often reacts bitterly when negative findings surface. A governor's highway safety representative from a northwestern state was obviously looking for endorsement rather than evaluation when he lamented, "I paid those people $40,000, and they couldn't even find one significant difference." The thought that his program was ineffective was inconceivable, and therefore he concluded that the evaluator must have been remiss.

This example is not an isolated case but a recurrent theme. Many administrators are happy to praise evaluation as an ally to their agency when it produces positive findings. When it produces negative findings, it is damnably inadequate.

The senior author of this paper has participated in a series of evaluation workshops which was designed to familiarize Governor's Highway Safety Representatives and their staffs, NHTSA personnel, and others with the rudiments of highway safety program evaluation. During the course of those workshops, one student showed a clear understanding of the concepts presented and seemed to have developed an appreciation for the process itself. However, at a subsequent meeting some months later this same individual presented the results of a well-conducted evaluation which he had overseen. The outcome was negative. But the individual concluded, "I am going to recommend that we continue the program anyway. I know the program is working -- in spite of the data." In this case, the individual knew how to carry out an evaluation, but was unable to accept the findings of an evaluation which proved contrary to preconceived notions.

The extent to which evaluations produce negative findings and are subsequently overriden by personal judgment is unknown. If an evaluation is unacceptable in terms of outcome, and if the individuals sponsoring or carrying out the evaluation reject the findings, then the evaluation will never be published. Undoubtedly, this phenomenon

occurs in the highway safety field. The prevalence of the phenomenon, however, is unknowable.

4. Resistance to measurement.

In order to initiate an evaluation, it is necessary that the goals of the project be presented in a clear and detailed manner. Furthermore, the goals of the project must be quantifiable. Any program which strives to "produce better drivers" or "improve driver attitudes" is not amenable to evaluation. These phrases are simply too vague and too qualitative to allow measurement to take place. Without measurement, evaluation is impossible.

Many program people, with the best of intent, do not feel that numbers capture the essence of what they are striving to accomplish in the name of safety. As one teacher from an ASAP rehabilitation clinic told the senior author, "I don't care what the evaluation shows, I know I am helping these people." In fact, the teacher did not know whether or not he was helping "these people," but it would be a difficult task to try to convince him otherwise. The very altruistic qualities which bring individuals into the field of highway safety often blind them to analytical assessments. With the conviction that they are "making things better," they forge ahead resenting any attempts to question or quantify the effects of their program.

A program sponsored by the Law Enforcement Assistance Administration (LEAA) in New York City is aimed at rehabilitating former convicts and assisting them in making the transition back into society. Concerning the effectiveness of the program, the director says:

> I just know in my gut that things are better because of the program, at least for some of the men. You just have to believe that things are better with us than without us (New York Times, March 10, 1974).

Indeed, things may be better because of the program, but declaration and conviction do not prove the case. Without hard data, the efficacy of the program is unknowable.

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science. (Lord Kelvin)

Summary.

Naively ignorant individuals often promote highway safety programs without any knowledge that their program should be evaluated. On other occasions, they call in "expert witnesses" to determine whether or not their program was of benefit. Occasionally, they even carry out more rigorous evaluations of their program in an effort to substantiate or "prove" that their program is of benefit. These people seriously damage the state of the art of highway safety evaluation, but they do so without malice and without the intent to deceive.

Other individuals who are far from naive have harmed the state of the art of highway safety evaluation to an even greater degree. These people are actively opposed to the execution of evaluations for reasons which are explained in the following section.

B. Administrator Wisdom

1. Anticipated negative results.

In fiscal year 1973, $109 million was appropriated under Section 402 of U.S. Code, Title 23 in order to bring the 50 states more in compliance with the 18 federal program standards. While $109 million is a large sum of money in absolute terms, it is a relatively small sum of money to spend in an effort to improve the safety requirements of a transportation system as massive as the one which exists in the U.S. The $107 million of 402 funding actually spent was divided, and re-divided, and re-divided again, until each state in the union had its share of the pie, until each state addressed each of the 18 standard areas, until one or more projects were carried out under each standard. The result of all the division of funds was to produce a series of projects and programs each of which was probably underfunded, and many of which were, as a result, ineffective.

If, for example, it were decided that a remedial driving course should be designed for purposes of improving accident rates of poor drivers, and if large sums of money were available to carry out extensive re-education and training with these drivers, then remedial driving courses might be shown to be of worth. But if funds are scarce and extensive retraining and re-education are not available, some 3 or 4-hour lecture series will be put together with minimal

dollars in order to satisfy the dictates of NHTSA. The program will almost certainly be fruitless.

Most valid evaluations have shown that driver education is ineffective (Joksch, 1972). However, it seems reasonable to expect that extensive training procedures could develop an individual to such a skill level that he would be capable of driving with a lower level of accident probability. Such a driving course would be very expensive and might be well beyond the finances of the nation. But the present "30 and 6" instructional package is little more than an exercise in futility. Human behavior is not easily changed. The notion that it can be dramatically altered in 36 hours, 30 hours of which are lecture, to produce a safer driver is ludicrous.

The above paragraphs essentially address themselves to a phenomenon which in medical/pharmaceutical parlance is referred to as a dose-response function. Most of the highway safety programs which are administered at the present time are woefully inadquate to accomplish their intended goals. If dosages could be dramatically increased, if remedial driving courses or driver education courses could be significantly strengthened, either or both of these programs might be shown to be effective. Present doses of driver education and remedial driving, however, may be so low that no benefit can be attributed to them. This phenomenon has been characterized by Campbell (1973) as dividing one aspirin among 10 people with a headache. Any study which seeks to determine the efficiency of aspirin in reducing headaches in such a situation would find them unbeneficial. Similarly, driver education and remedial driving may be unbeneficial not because the concept is untenable, but because the applied doses have been too small to produce measureable effects.

Many state administrators and many program managers are aware of the fact that they are overseeing projects which are underfunded and watered down to such a level that they cannot be effective. Many of these administrators continue to work with their programs in the hope that they will serve as pilots to other projects, in the hope that methods can be improved, and in the hope that some day more funding might be available to increase their program dosage.

Not surprisingly, these administrators actively resist evaluation of their programs. They know that their programs are inadequate, and they know that their programs, when subjected to rigorous evaluation, will be found wanting. To allow their program to be evaluated, then, seems to be injurious to their best interest and to any hopes which they might have for future program development. Therefore, whenever evaluations are required of these administrators, they may act to

avoid and/or sabotage the evaluation process in whatever ways they find available.

2. Insensitive dependent variables.

Many administrators anticipate that an evaluation of their program will result in negative findings due not only to the weakness of the program itself, but due to the insensitivity of the variables used to measure program effectiveness. Death rates, injury rates, and accident rates are the legitimate measures of highway safety program benefit. Each of these measures, however, may well be insensitive to potential benefits of a given program for two basic reasons: (a) Accidents, particularly fatal and severe accidents, are very rare events. The motor vehicle death rate in the U.S. is slightly over 4 people killed per 100 million miles travelled. The driver of a passenger car is killed in only one accident in 200. Logically, then, any program which seeks to reduce death rate can be determined effective only after large numbers of measurements are taken. If a program were only moderately successful in reducing driver fatality rate, thousands of crash-involved vehicles exposed to the program and not exposed to the program would have to be monitored in order to make mathematically rigorous judgments concerning the program. (b) Motor vehicle accidents are unusually complex events. They are the resultant of malfunction, misjudgments, incorrect responses, errors of perception, visual deficits, and thousands of other things. Any program which is set up to improve driver behavior, to reduce vehicular speeds, or to get the drunk off of the highway, addresses only one, or at best a few of those dimensions which are associated with and account for accidents. The almost capricious nature of accidents causes death rate, injury rate, and accident rate to be insensitive to the goals which the program seeks, in the short run, to redress. For this reason, when administrators and program managers are prevailed upon to evaluate their program, they frequently seek to use a dependent variable (a proxy measure) which may be more responsive to the program which they have constructed. For example, in evaluating a remedial driving course, the program manager would prefer to evaluate his course on the basis of material learned, attitudes changed, etc. A driver education teacher might well resist having his or her students' accident rates compared to a suitable control group, but they would probably be agreeable, if not downright anxious, to have their students compared to a control with respect to knowledge of rules of the road, shapes of road signs, and the minimal insurance requirements for vehicle operation.

To reiterate, highway safety administrators and program managers are often well aware of the fact that their programs are underfunded

and of such modest import that they cannot possibly have large-scale effects on the legitimate measures of highway safety benefit. They realize that the legitimate dependent measures of highway safety are insensitive dependent variables by which to gauge a program. If they are forced to evaluate their own performance, or if they are forced to accept the evaluations of others, they usually endeavor to substitute a proxy measure in the evaluation design which is more likely to show positive program accomplishment. Better still, if at all possible, these people will avoid taking part in an evaluation altogether, fending off calls for evaluation with such phrases as "we don't have time to do what would be required for a proper evaluation," or "we didn't know that this project was going to be evaluated and, therefore, we didn't collect appropriate measurements at the inception of the program."

3. "Over-advocacy"

> It is so difficult to get our sluggish bureaucracies to make any change that advocates have to promise more than they can really deliver to sell a real program. This means that a realistic evaluation of the new program is political suicide for those who advocated it (Salasin, 1973, p. 9).

Aggravating the phenomenon of anticipated negative results is the phenomenon of over-advocacy. In order for a program manager to secure funds for his project, it is necessary that he convince some agency or legislature that what he is endeavoring to accomplish via his program will result in highway safety benefits. Furthermore, because of the built-in inertia of most agencies and legislatures, it has become necessary for program advocates to "sell their program," when they request funds. Managers and administrators are in competition with others for a limited number of dollars. Since the competition is often keen, it stands to reason that those programs which will produce the most benefit or, at a minimum, promise to produce the most benefit stand to be first in line for funding. Such an allocation procedure results in over-statement by program advocates to the point of downright falsehood. Managers promise results which they cannot possibly deliver. Administrators claim that they will provide benefits which are unattainable. In the long run, these individuals are found out. Their glittering generalities are no longer accepted. In the short run, however, these same individuals strive to avoid detection. They work to prevent any evaluation of their project since performance will almost certainly fall short of promise.

III. INTERNAL IMPEDIMENTS TO PROGRAM EVALUATION

A. Lack of Technical Knowledge

There are numerous internal impediments which interfere with the successful performance of the various steps in the evaluation process once it has been initiated. Many impediments which occur in evaluations are the result of poor judgment or lack of knowledge on the part of the evaluator. Six such impediments to the successful completion of evaluations are: (1) program goals too broadly defined, (2) incorrect choice of experimental design, (3) inappropriate use of proxy measures, (4) improper feedback of information to administrators, (5) capitalization on regression toward the mean, and (6) inappropriate choice of statistics.

1. Program goals too broadly defined.

As has been stated several times, there are three goals in highway safety: (1) reduction in loss of life, (2) reduction in injuries, and (3) reduction in property damage. It does not follow from this, however, that all highway safety countermeasures should be effective in all accident situations. For example, if an evaluator seeks to determine the effectiveness of median barriers, he would obviously be interested in accident rates at various sections of highway which contain or do not contain median barriers. But would the evaluator be interested in all accidents which occur at the median and non-median sections of the highway? Perhaps not. The main purpose of median barriers is to prevent vehicles from crossing over the center line and striking other vehicles head on. Therefore, the evaluator might reasonably decide to look at the rates of head-on accidents at comparable sections of highway both equipped and not equippped with median barriers.

As another example, consider how an evaluator might go about determining the effectiveness of reflectorized license plates. Obviously reflectorized license plates have no effect on the rate at which vehicles are struck in the side. On the other hand, if a reflectorized license plate enhances the visibility of a vehicle when seen from the rear, then the rate of rear-end collisions for cars equipped with reflectorized plates may be less than the rate for cars not so equipped. But the enhanced visibility of reflectorized plates is evident only after dark -- reflectorized plates are no more visible than ordinary plates during the daylight. In short

then, the evaluator of reflectorized plates might well confine his evaluation to rear-end accidents occurring after sundown (see Stoke, 1974).

By narrowly defining those accidents which a given program or countermeasure is designed to reduce, the evaluator gives the program the maximal opportunity to show its effect. By broadly defining what the program attempts to accomplish, accidents which the program was not designed to affect may enter the data, thereby watering down and destroying what might otherwise have been a significant effect.

2. Incorrect choice of experimental design.

The six research designs most often used in conducting highway safety evaluations are: (A) after-the-fact design, (b) before-after design, (c) before-during-after design, (d) time series analysis, (e) before-after design with control group(s), and (f) a posteriori design. All of these designs have been and are being used, some producing fallacious results. The following is a description of the various designs, with mention of the strengths and weaknesses of each.

(a) After-the-fact design.

"Since our program was put into effect, only 50 people have been killed on the city's streets." This glib phrase is exemplary of what might be called a declaratory design or an after-the-fact design, or, more appropriately, an after-the-fact fallacy. By the tenets of this design, a declaratory statement is made concerning the effectiveness of a particular program. The statement tends to give the impression that: (1) after the introduction of a specific program, a given state of affairs (e.g., death rate or accident rate) improved, and (2) the program was responsible for the improvement. This design does not directly measure death rate, accident rate, or any other dependent variable before initiation of the program. Instead, it simply alludes to a poorer state of affairs which existed before the program was undertaken. Obviously, if the rate of deaths or accidents before the creation of the program is unknown, the effectiveness of the program cannot be calibrated.

(b) Before-after design.

It is hoped that administrators and the public in general would not fall prey to the after-the-fact design. Because the fallacies inherent in that design appear to be so obvious, it seems that it would be accepted by only the most naive of judges. Indeed, such a declaratory design seems to be on the decline. The before-after

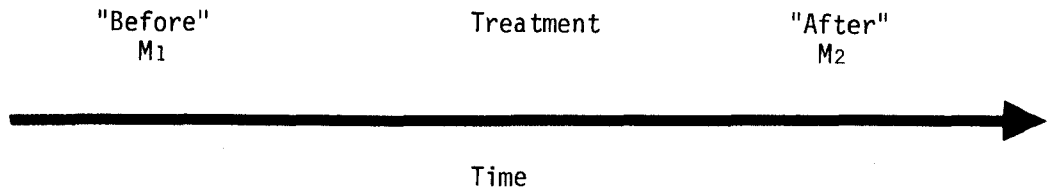
design, however, still has many supporters, and the frequency with which it is used in evaluations seems to be on the rise.

The before-after design is a relatively straightforward procedure calling for two measurements to be taken: one before a program is instrumented, and one at some time after the program has been in progress. Program effectiveness is defined as the difference in these two measurements across time (see Figure 2).

Those people who quickly reject the after-the-fact design as being basically fallacious sometimes have more difficulty seeing the fallacy in this design. But the fallacy nonetheless exists. Logicians insist that this before-after design commits a post hoc ergo propter hoc (after-the-fact therefore because of the fact) fallacy. Two implicit assumptions underlie the before-after design: (1) if the treatment (program) had not been instrumented, the measurement taken before treatment was introduced would have continued at the same level into the indefinite future, and (2) if an improvement is seen in the after measurement, the improvement is attributable to the program. Both of these assumptions can obviously be wrong. If either or both are wrong, the design may yield an invalid outcome.

(c) Before-during-after design.

This design attempts to overcome limitations of the before-after design. By means of this design, measurements are taken at three points in time -- before the program is put into operation, while the program is in operation, and after the program is terminated (see Figure 3). If the program is effective, the second measurement (the one taken during the program) will be shown to be associated with the lowest death rate, accident rate, etc. And ideally, death rate, accident rate, or whatever variable is monitored should be approximately equal for measurements one and three, before and after the program. When this idealized state of affairs does not materialize, interpretation of the data becomes more difficult. For example, if death rate is high before a program is instrumented but low during the program and low after the program is terminated, the effectiveness of the program is in some doubt. It could be submitted either: (1) that death rate was generally declining and the program had little or no effect on that decline, or (2) the program had a direct effect while it was in operation and a residual effect after it was terminated. Either is possible. (For additional information on this design, see Sidman, 1960.)



Effect = M_1-M_2

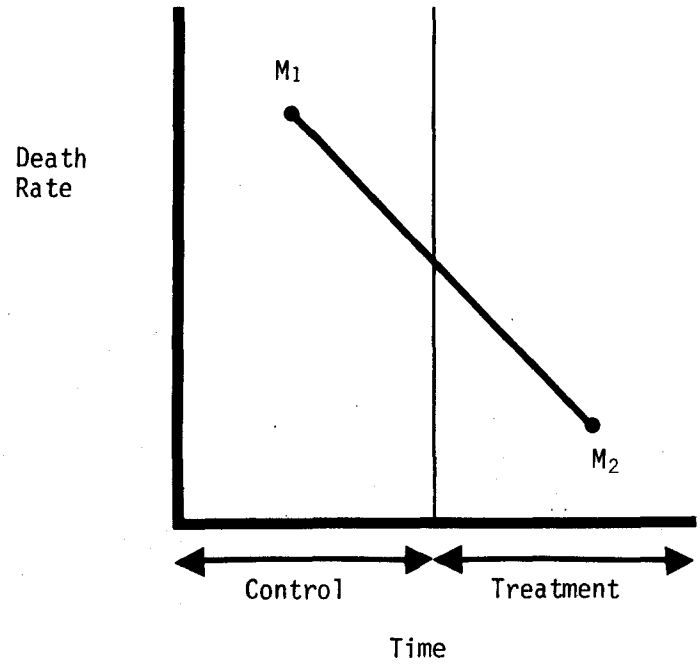


Figure 2. Before-after design.

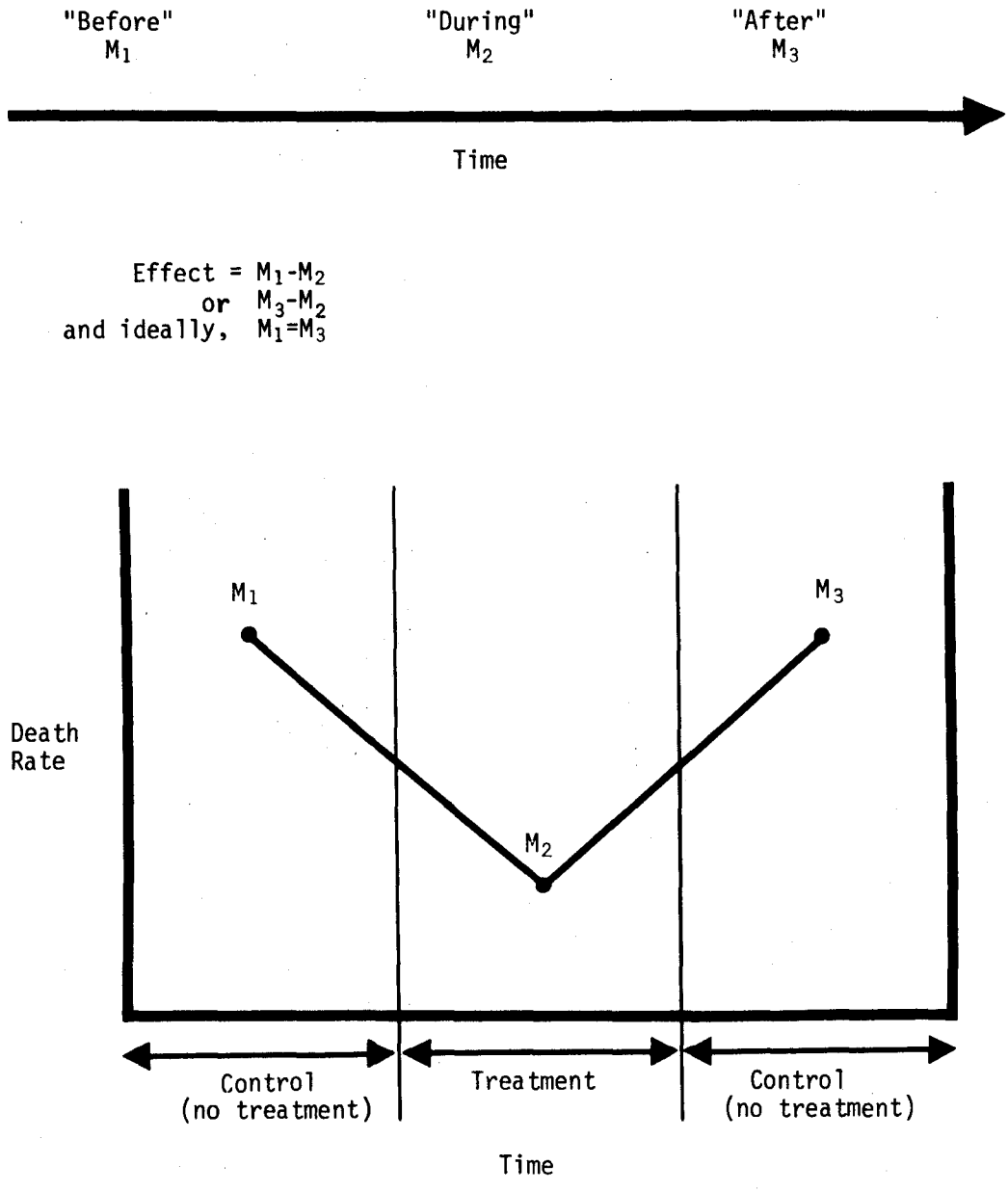


Figure 3. Before-during-after design.

The main use of the before-during-after design in the field of highway safety has been in the general area of enforcement. The reason for this is quite simple. Most highway safety programs cannot be readily terminated for purposes of evaluation; many enforcement programs can. While it is not feasible to put up and take down guardrails, it is quite possible to double police enforcement on a given section of highway and then return to the normal enforcement procedures. By so doing, a reasonable measure of enforcement benefit can be calculated.

(d) Time series analysis.

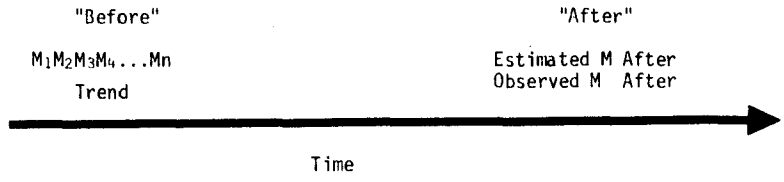
This design attempts to overcome the first tacit assumption in the before-after design (i.e., if the treatment program had not been implemented, the before measurement would have continued at the same level into the future). By taking multiple before measurements, users of time series analysis attempt to discern any underlying trends, across time, in death rate (or accident rate, or whatever variable in being monitored) which are independent of the treatment program (see Figure 4).

While this design overcomes the first basic flaw in the before-after design, it does not necessarily overcome the second -- that is, if an improvement is seen between before and after measurements, it is attributable to the treatment. Any condition which covaries with the introduction of a program can at least partially account for the program's "effectiveness." As a case in point, in the winter of 1973-74, the nation changed over to a 55 mile per hour national speed limit. Simultaneously, the national death rate, which had been slowly declining over the past several years, took a decided plunge. Why? Perhaps because people were indeed driving slower. But any other set of circumstances which was dramatically altered during that period (e.g., driving patterns) could at least partially account for the results. (For more information on this design see Ross, 1974; D. T. Campbell, 1972.)

(e) Before-after design with control group(s).

This design is most familiar to those trained in pure experimental research methodology. It is the most powerful of the six designs discussed herein. Unfortunately, in the field of highway safety, this design is probably used less than any of the other five.

By the dictates of this design, two (or more) comparable groups are set aside for comparison. Prior to treatment, measurements are taken of both groups. One group, the experimental group



Effect = Estimated M_{After} - Observed M_{After}

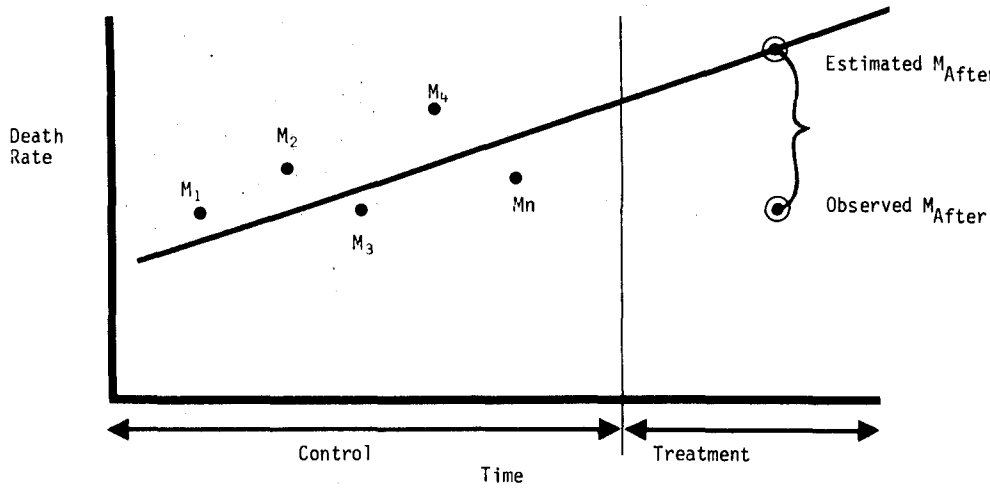


Figure 4. Time series design.

(treatment group), is exposed to some highway safety program. The other group (control group) is not exposed to the program. After some time has elapsed, both groups are measured to determine death rate, accident rate, etc. The difference between the two groups is the measure of the program's effect (see Figure 5).

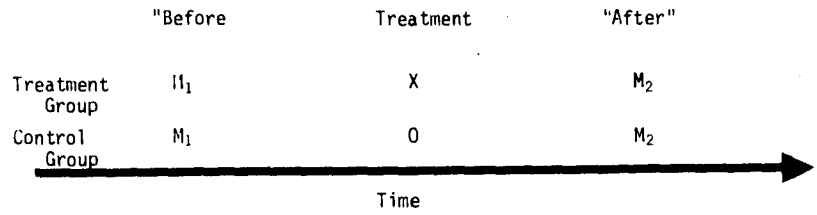
(f) A posteriori designs.

A posteriori designs are very similar to the standard experimental design with control groups, with one basic difference. In the experimental design, the evaluator chooses comparable groups (cities, drivers, sections of highway), makes a manipulation on one, and saves the other for a control group. In a posteriori designs the treatment group and control group are not determined by the evaluator, but by real-world circumstances.

If an evaluator wanted to determine the effectiveness of seat belts in reducing fatalities via standard experimental procedures, he would dictate that one group of people always wear belts while a comparable group always go beltless. After both groups had been involved in numerous accidents, he would calculate the fatality rates for each group. By means of these calculations, the effectiveness of belts would be determined.

Obviously, such an experiment would be inappropriate monetarily, temporally, and ethically. An alternative is to use an a posteriori design. This design capitalizes on the fact that many people ride in automobiles while wearing belts, and others ride beltless. To the extent that the two groups have different fatality rates, it might seem reasonable to assume that the effectiveness of the belt could be calculated in a straightforward manner. But it must be remembered that the two groups (belted and non-belted) were not randomly assigned by an evaluator. Therefore, there is no reason to believe that both groups were necessarily comparable to begin with. And if, for example, belt wearers tended to be middle-aged drivers, driving newer cars, at lower speeds, etc., then some of the effectiveness of the belts might reasonably be attributed to these other variables. In fact a posteriori designs recognize that these covariants can and do interact with the main variable of interest. By appropriate mathematical weighing procedures, this design attempts to subtract out the effects of certain specified covariants. (For examples of this procedure see Tourin & Garret, 1960; Campbell, 1968, 1971, 1974).

Once the design of an evaluation is established, measurements can be taken. The measurement phase of an evaluation is the most



Effect = M_2 (treatment) - M_2 (Control)

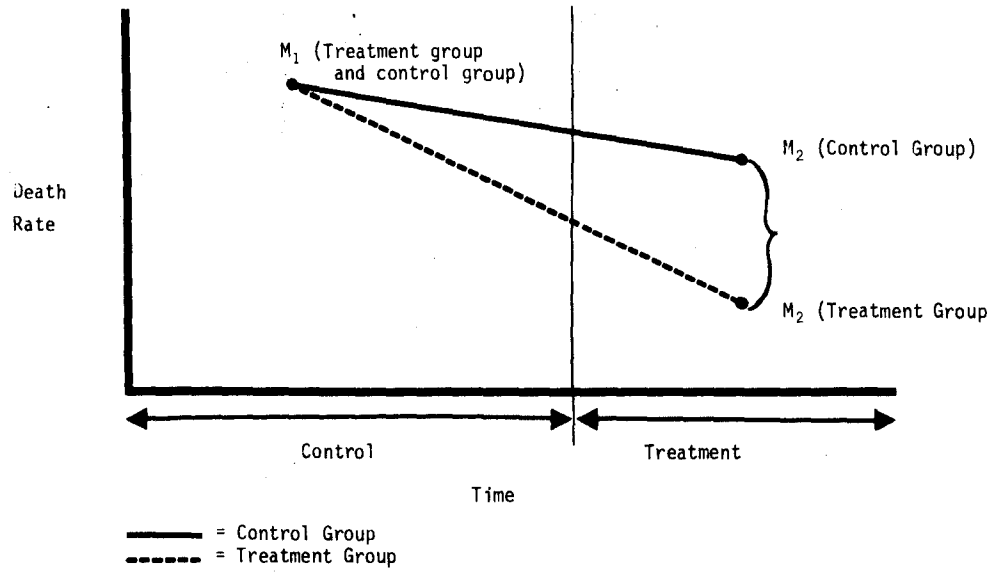


Figure 5. Before-after design with control group(s).

time-consuming, expensive, and important part of the whole evaluation process. If the evaluator overcomes the various obstacles to relevant and valid measurements, he has traveled over half the distance to a sound conclusion.

When evaluations are attacked, they are usually attacked on the grounds of inadequate design or inappropriate statistical techniques. While design errors and statistical errors can certainly render an evaluation invalid, it is more frequently the case that evaluations err because of poor measurement. Without accurate measurement, any design is spoiled; without accurate measurement, all statistical inferences are without basis. Without valid inferences, the conclusions of any evaluation are fallacious.

Summary.

Of the six evaluative designs just mentioned, some were seen to be valid and appropriate, while others were deemed invalid and erroneous. It should be noted, however, that the validity or invalidity of these various designs were determined by logical, rather than practical, means. A very practical issue which must necessarily be addressed is the applicability of all six designs in real-world settings. The question must be asked: "Will true experimental designs work with human beings, when the laboratory is the community, and when the stimulus is a new social program?" (Kershaw, 1972, p. 20).

Weiss and Rein (1969) state:

> ...communities are open to all sorts of idiosyncratic experiences from the personality of mayors through the location decisions of industries. What the comparison 'sample' really accomplishes, from a statistical point of view, is that a single case in which there is no intervention is being compared with a single case in which there is an intervention. The statistical merit of this procedure is very close to zero (p. 140).

Weiss and Rein paint a very pessimistic picture of classic experimental procedures as they are applied to real-world evaluations. In effect, they question the validity of experimental procedures outside of the laboratory, not because of any deficiency in experimental procedure per se, but because of the hostile environment in which the procedures must be conducted.

While it should be readily acknowledged that experimental research pursued outside the antiseptic environs of a laboratory is open to numerous contaminants, it should not be concluded that use of this procedure should be abandoned. The exigencies of the real world may force investigators to define variables with less precision than would be desirable; the dictates of society may prevent the evaluator from exercising as much control as would be wished; data collection mechanisms might be less than ideal; but these differences in real-world procedures and laboratory procedures are quantitative and not qualitative.

3. Inappropriate use of proxy measures.

In the United States today, there are over 125,000 registered motor vehicles. Each year these vehicles travel over 1.25 trillion miles and in the process kill 50,000 people, while injuring many times that number. Though the absolute number of people who are killed and injured in motor vehicle accidents is large, the rate of death and injury is low. From this it follows that motor vehicle death rate and motor vehicle injury rate are insensitive dependent variables with which to measure highway safety program effectiveness.

If a safety program or countermeasure is shown to be effective in reducing fatality rate or injury rate, either:

(a) the program itself is extremely beneficial and can be so demonstrated with limited amounts of data, or

(b) the program is of modest benefit and can be shown to be effective only by collecting large masses of treatment and control data over extended periods of time.

Most highway safety programs are rather modest in scope. Few administrators of these programs feel that their efforts will eliminate highway deaths or even drastically reduce death rate. Therefore, if their programs are to be demonstrated effective in reducing deaths, they must do so by following the dictates of the second proposition above.

For some programs, "it would take years, even decades, to test the program's effectiveness in achieving its long-range expectations. In the interim, proxy measures have to be used that are germane to more immediate goals and presumably linked to desired ultimate outcomes" (Weiss, 1972, p. 37).

Proxy measures are variables which can be readily collected and which are assumed to be correlated with accident avoidance or accident attenuation. Thus, if high speeds are associated with an increased probability of an accident, and if a given countermeasure reduces vehicular speeds, it seems likely that the countermeasure will also be effective in reducing death, injury, and property damage. If the wearing of seat belts can be shown to correlate with reduced injury and death, and if a given program results in increased seat belt usage, it follows that the program should be efficacious in reducing deaths and injuries.

Unfortunately, many proxy measures are used in highway safety evaluations which have not been shown to correlate with death, injury, and damage. Foremost among these invalid proxy measures are "public awareness" and "attitude change."

In the early 1970's, the Texas Traffic Safety Administration was actively engaged in a "Drive Friendly" campaign. This campaign was carried to people by way of television announcements, road signs, bumper stickers, and billboards. The intent of the campaign was to promote driver courtesy and improve driver attitudes. Subsequently, the question was asked: "Is the public aware of the program?" A large scale survey was undertaken to answer the question. When the results were in, it appeared that indeed the public had gotten the message. But at this point a more relevant question should have been asked: "Is there any correlation between public awareness of safety campaigns and accident rates?" Only if this question can be answered in the affirmative is the proxy measure "awareness" appropriate. If not, the whole evaluation was an exercise in futility.

In a study carried out by Malfetti and Simon (1974), the effect of a "DWI-Counterattack" program was evaluated. The authors describe the program as one which, "substitutes re-education and rehabilitation for traditional punitive measures since punishment alone does not seem to work" (p. 50). The program was offered to those persons who had been convicted of alcohol-related violations or had refused the chemical test for blood alcohol content. The course consisted of five, two-and-one-half hour sessions conducted by the professional staffs of nearby educational institutions.

In describing the program and its evaluation, the authors stated:

> The ultimate objective of the Westchester course is behavioral modification of the students, specifically in reduction or elimination of their DWI habits. A definitive evaluation of the course in terms of this objective

> would require continuous monitoring of the actual drinking and driving behavior of large numbers of course graduates (and of a control group of comparable persons who did not take the course) for several years within the framework of an appropriate experimental design. Obviously, such direct measurement is impractical and such a definitive evaluation cannot be made. In lieu of this, various indirect measures were considered for an evaluation of the course (p. 51).

They further explained that subsequent DWI citations and convictions were not used to determine program effectiveness because of difficulties in validity and reliability; loss, misplacement, and inaccuracies in the records themselves; and the costliness and time-consuming quality of data collection. Thus,

> Levels of knowledge and attitude toward alcohol and driving were selected by the authors as measures of change produced by the course.... it was judged reasonable to investigate the extent to which the course was successful in increasing relevant information and improving attitudes, and to assume that persons educated to possess accurate information about the effects of alcohol and the implications of their own drinking and driving behavior would be in a better position to adopt appropriate countermeasures. Moreover, 'before' and 'after' measures of knowledge and attitude could readily be obtained as part of evaluation measures already built into the course, and could be tabulated and analyzed easily and inexpensively (p. 52).

Some would argue that using "public awareness and attitude change" as proxy measures may be the best indicators presently available with which to evaluate certain programs. Accordingly, they would argue that even though the correlation between these variables and the legitimate aims of highway safety is negligible, in the absence of other, more appropriate proxy measures, these measures should be used. This is not necessarily true. By using proxy measures which are inappropriate, a false sense of security is created. If a proxy measure indicates that a program is effective, the program would probably not be shut down, and in fact it might be expanded. But if the proxy measure does not correlate with death, injury, or damage, this recommendation is without foundation. Furthermore, by continuing to use inappropriate proxy measures, no progress toward the establishment of more meaningful measures is made.

Kaestner (1974) addresses the question of using attitudinal measures to assess driver improvement programs. He says:

> A review of the literature by this writer uncovered innumerable inferences 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 commensurate 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 (p. 6).

If the only choice is between evaluating with a proxy measure having little or no correlation to the legitimate aims of highway safety, and not evaluating the program at all, the latter should be chosen.

4. Improper feedback of information to administrators.

Some administrators prevail upon evaluators to feed back preliminary evaluative information so that they might make "improvements" in the program while it is in progress. Such feedback, some proponents argue, is not only expedient, but ethically required. Brooks (1971), for one, argues in favor of "the ethical necessity for continuous feedback of research findings into community action programs, thereby producing adjustments and improvements in their operation" (p. 37). By taking this position, he realizes that he does harm to the original research design.

> While this is the correct procedure from the ethical and action point of view, it has the unfortunate effect of tossing a monkey-wrench into the research design constructed at the program's outset. The person interested solely in the research implications of a program might prefer that it be carried out through to completion without alteration, whether successful or not, so as to yield unsullied findings of maximal generalizability. Given the social ethic which underlies the community

> action program, however, it is necessary to devise an evaluation procedure which not only accommodates, but in fact facilitates the feedback process (Brooks, 1971, p. 37).

In effect, Brooks takes the argument of ethics and twists it to prevent any possible evaluation of those programs deemed to be community-action oriented, or politically sensitive. This is a negation of evaluation and a negation of the ultimate responsibility of evaluators to determine the efficacy of social action. Unnecessary feedback or leaks make a mockery of the evaluation process. Unless the program is causing harm (a relatively rare occurrence in the field of highway safety), it is more ethical to prevent feedback and to insist that the program remain stable for a reasonable period of time so that the simple question, "does it work?" can be answered.

> The ethical investigator protects participants from physical and mental discomfort, harm, and danger ...A research procedure may not be used if it is likely to cause serious and lasting harm to the participants...where research procedures may result in undesirable consequences for the participant, the investigator has the responsibility to detect and remove or correct these consequences... (American Psychological Association, 1973, p. 2).

If an evaluator feels that a program is injurious, obviously he is obliged to say so. But there is nothing pious or commendable about feedback which results in spasmodic shifts of program personnel or program emphasis. Feedback, per se, can be an impediment to the whole evaluation process. It is wise for the evaluator and the administrator to agree in the beginning how long the program will remain constant, without experiencing shifts and modifications which guarantee to make the original research design obsolete and useless.

5. Capitalization on regression toward the mean.

> On December 23, 1955, Connecticut instituted an exceptionally severe and prolonged crackdown on speeding. Like most public reporting of program effectiveness, the results were reported in terms of simple before-after measure: a comparison of this year's figures with those of a year ago. The 1956 total of 284 traffic deaths was compared with the 1955 total of 324, and the governor stated, "With a savings of 40 lives in 1956...we can say

the program is definitely worthwhile.' (D. T. Campbell, 1972, p. 121).

The citation above is a classic example of the phenomenon of regression toward the mean. First recognized by Francis Galton in the late 1800's, this phenomenon has probably accounted for more "program successes" than any other fallacy perpetrated upon the highway safety community. Regression toward the mean is a mathematical phenomenon which , simply stated, says that if two measures are associated with less than perfect correlation, unusually high or low scores on one measure will tend to be associated with more average (mean) scores on the second. If the number of traffic violations which people commit one year are minimally or moderately associated with the number of traffic violations they commit the second year, and if a given individual commits an exorbitant number of violations one year, it should be predicted that he will commit a more average number of violations the next year. Similarly, if an individual is free of violations one year, the best guess of the number of violations he will commit the next year is a number above zero and less than average.

Assume that the number of accidents sustained at an intersection one year is only modestly associated with the number of accidents that will be sustained at that intersection the next year. Now, further assume that the intersection has witnessed an unusually high number of accidents this year. How many accidents will occur at this intersectio next year? The best guess is a number less than occurred this year, but more than average. Even if no attempt is made to improve the intersection this year, even if no new crosswalks are installed, no lights or signs set in place, no additional enforcement personnel assigned, a reduced number of accidents would be expected at this intersection next year. From all of this it follows, that if crosswalks, lights, signs, or enforcement personnel are added to our bad intersection, it is <u>not</u> proper to conclude that the difference can be accounted for by the treatment imposed. Indeed, a reduced number of accidents would have been expected had nothing at all been done.

It should be noted that the lower the correlation between two measures, the more salient is the phenomenon of regression toward the mean. As a limiting case, if there is zero correlation between two variables, regardless of the score on the first variable, the best guess of the associated score on the second variable is the average (mean). If there is no correlation between the number of accidents an individual has one year and the number of accidents he has the next year, then the fact that Mr. Jones had eight accidents last year is in no way indicative of how many accidents he will have

this year. Similarly, if accident rates have only a very low correlation from year to year, the knowledge that Mr. Jones was involved in eight accidents last year moves our best guess of his accident record for this year only slightly away from the average (mean) and toward eight.

The two sub-areas in highway safety which have been most burdened with the regression toward the mean fallacy are: (1) driver improvement programs and (2) highway improvement programs. Numerous studies in these two areas are found throughout the literature. And in each study the same pattern repeats itself: (1) an unusually bad "before" sample is collected (e.g., a group of drivers who have had four or more violations in a year; a section of highway which had an inordinate number of ran-off-road accidents in a given year), (2) some treatment program is initiated (e.g., the bad drivers are given a driver rehabilitation course; the highway is widened), and (3) the "after" data show an improvement "due to treatment" (e.g., the bad drivers have fewer violations the next year; the widened highway is associated with a lower frequency of ran-off-road accidents).

All before-after evaluations are highly suspect, logically. All before-after designs which contain unusually egregious before data should be discarded without further consideration.

6. Inappropriate choice of statistics.

> If, too often we find that figures fool, it is because too often fools figure. (J. P. Guilford, 1936)

If a program evaluation has been well designed, if good data have been collected, and if the treatment group seems to differ from the control group, inferential statistics can be employed to see if a "significant" difference exists. Which inferential statistic should be employed? Chi-square? Analysis of variance? Analysis of covariance? Without knowing what type of data were collected and without knowing the conditions (design) by which the data were collected, this question cannot be answered.

While it is sometimes difficult to state which inferential statistic should be chosen for a given evaluation, the following example shows how not to choose an inferential statistic. At a meeting attended by the senior author, several administrators from the southeastern United States were gathered together to discuss some mutual problems which they were having in evaluating a particular project. A representative of the National Highway Traffic Safety

Administration (NHTSA) was present to advise and assist the administrators in trying to resolve their difficulties. Near the end of the meeting, one of the administrators asked the NHTSA representative which statistical procedure they should use in analyzing their data. He responded, "analysis of covariance." When asked why, he said, "Because it will show the most significant benefit."

Whether or not analysis of covariance would have shown "the most significant benefit" is not an issue. Whether or not analysis of covariance was the appropriate procedure to apply in this case is not questioned. What is questioned is the reason why the procedure was chosen. Inferential statistics should be chosen on the basis of data and design, not desired outcomes.

B. Inadequate Tools, Procedures, Data Bases

There are several internal impediments to evaluation which are the result of real-world, limiting situations. These are: (1) changing program goals, (2) inappropriate data collection methods and data bases, (3) poorly timed phase-in procedures, and (4) lack of control groups.

In the following pages, these impediments are discussed.

1. Changing program goals.

In some cases, program people specify their goals in clear quantitative terms at the outset of the project, and then proceed to change their goals during the implementation of the project. This changing of program goals poses a real problem for the evaluator, and often it is a problem of which he is unaware.

One ASAP program developed a remedial program for DUI offenders who were classified as "social drinkers." It was felt that social drinkers (defined by blood alcohol levels just above the legal limit) were amenable to persuasion aimed at encouraging them not to drive after they had had a certain number of drinks. While the proponents of the program conceded that the course would be of little benefit to a chronic alcoholic, it was hoped that reminding the social drinker of his responsibilities while driving and impressing him with the hazards of drunk driving might thereby enhance his behavior.

After the course was designed and funded, DUI offenders with marginal blood alcohol levels were enrolled. Ideally, the offenders who completed the course would have been compared to similar

individuals in the control group -- comparable offenders who had not attended the school. Unfortunately, during the course, the teacher became deeply troubled by the fact that so many DUI offenders were missing out on the opportunity to sit through his class. While conceding that the program was not designed for the chronic drunk, the teacher insisted: "As long as there is the possibility that some of these people might benefit from the course, and as long as I have room left in my classroom, they should attend." Accordingly, arrangements were made whereby the more chronic drinkers were enrolled.

By allowing chronic offenders to attend the remedial course, the nature of the experimental group was changed -- and changed in a direction that would make subsequent evaluation underestimate the effectiveness in the course.

In terms of this particular example, it was probably inappropriate to allow the chronic drunks to attend the course. Since the program was not designed to rehabilitate this group, and since efforts spent working with these students were efforts subtracted from others, it seems reasonable to conclude that the admission of this group was ill conceived, altruism notwithstanding. But, were it deemed necessary that these more serious offenders be admitted to the course, the evaluator should have been informed immediately. Had he been informed, some coding procedure which distinguished the social drinkers from the chronic drinkers could have been devised. Then, when comparisons were made between the experimental group and the control group, the chronic offenders who attended the class could have been separated out. By separating out this group, a better evaluation of the effectiveness of the course with respect to its stated goals could have been performed.

When reading through published evaluations, it is difficult to know whether or not this "error of program change" has been committed. As was suggested earlier, programs often undergo change during development and implementation which are not reported to the evaluator. Not realizing that procedures, definitions, or experimental and control groups have been altered, the evaluator carries out his tasks under the original set of assumptions and thereby produces fallacy.

In order to reduce the likelihood of this type of error, the evaluator should be admonished to stay close to the data. While it is possible that subtle changes in program inflection and intent can take place without the evaluator's knowledge, these changes are relatively less likely to occur if the evaluator stays in close touch with the program. And if it is decided that changes must be made in

the program, the evaluator is much better off if he knows about the changes from the very first.

2. Inappropriate data collection methods and data bases.

The fact that highway safety evaluations are not carried out in laboratories makes the problem of data collection much more difficult for evaluators than for traditional researchers. Instead of collecting the data himself or with the aid of one or two trained assistants, the program evaluator must work with data provided to him by highway patrolmen, local police officers, hospitals, etc. Often the quality of these data is poor due to different assumptions and definitions used by the various data collecting agencies. On other occasions, data have been shown to be completely fallacious, due to misunderstandings on the part of the people collecting the data, or the evaluator requesting the data.

One ASAP program sought to determine the effectiveness of a hospital clinic in improving the recidivism rate of DUI offenders. By enlisting the aid of local judges, it was arranged that one group of DUI offenders would be sent to the hospital clinic for treatment while a comparable group of offenders would be used as a control. The effectiveness of the program would be determined by examining the average amount of time which elapsed before the offenders were arrested again for DUI. Presumably, if assignments to the treatment group and control gruop were random, and if the treatment group went for longer periods of time without rearrest for DUI, the program was effective.

Unfortunately, close tabs were not kept on the treatment group. And, the fact that a judge sentenced an individual to a hospital clinic did not necessarily mean that he attended the clinic. On the first night the clinic was held, the roll was called and some individuals were found to be absent. On the second night the clinic was held, the roll was called again and some individuals sentenced by the court to attend the clinic still had not shown up. On the third night, if an individual again failed to answer the roll, hospital personnel assumed that the individual had no intention of attending the clinic. At this point they labeled the individual as "discharged" -- even though he had never attended a single session -- and sent his name to ASAP personnel. Accordingly, "discharged" individuals who never attended the clinic were placed in the treatment group along with "discharged" individuals who went to each session!

How often does this type of error occur in the literature? There is no way of knowing. Not that reputable evaluators would purposely submit bad data; rather, evaluators may be unaware that their own data are bad. Unfortunately, there is no certain cure for this malady.

3. Poorly timed phase-in procedures.

As has been suggested several times before, evaluation is often an afterthought to the development of a highway safety program. Frequently, programs are devised, funded, and staffed before any notion of evaluation of the program's effect has ever occurred to anyone, including the program manager. Without considering the need to evaluate the program, the program is launched and the opportunity to carry out a rigorous evaluation is lost.

The phasing-in of a highway safety program without consulting an evaluator can, and often does, prevent an evaluation of the project from ever being carried out. Inappropriate phase-in procedures thus can be thought of as an external impediment to the evaluation process itself. More often, however, highway safety programs are initiated and subsequently subjected to less than rigorous, makeshift evaluations. Because the evaluator is not a party to the project at outset, he is forced to undertake an evaluation under less than ideal circumstances. Control groups which could have been established at the outset of a project are left unestablished. Baseline measures of accident rate, fatality rate, or injury rate before the experimental treatment was introduced are not collected. Individuals are assigned to treatment or control groups for inappropriate, often biasing, reasons.

In short, the presence of an evaluation at the outset of a project does not insure that the project will be correctly evaluated. But if the evaluator is brought into the project only after substantial fundamental procedures have been established, it seems likely that the resulting evaluation will be of a lesser quality than it might have been.

4. Lack of control groups.

In an earlier section of this paper, six different experimental and quasi-experimental designs for carrying out evaluations were discussed. Some designs were shown to be fallacious, and some designs were shown to be better than other designs. The most powerful of the designs discussed, however, was the before-after design with control group(s). As the name of this design implies, the benefits derived from a program are ascertained by comparison to a control group.

For example, if one were interested in determining whether or not a medical review board was effective in identifying and treating medically impaired drivers (e.g., confiscating their licenses, restricting the time of day in which they could drive, only allowing them to drive with another person in the car), one would define a pool of drivers with potential impairments and randomly choose half of them to go before a medical board to receive whatever disposition the board found appropriate. The other half of the potentially impaired pool of drivers would serve as a control. At some future point in time, the accident records for the two groups would be compared. Since the groups were randomly chosen (control vs. medical review), any lower accident probability associated with the group which underwent medical review might logically be thought to have derived from the actions of the medical review board.

At this point it might be asked: "Is the experimental procedure defined above ethical? Can we ethically defend the notion of allowing drivers subject to coronaries access to the road? Can we allow drivers with poor hearing, poor visual acuity, or reduced visual field to operate their vehicles on the same roadways with more 'normal' individuals?" The answer to these questions is: if we do not know a priori whether or not a given treatment will reduce deaths, injuries, etc., then we are under no obligation to forego a control group. If, on the other hand, we do believe and have evidence to demonstrate that a treatment would be beneficial, then that treatment cannot be denied groups which are in need of it solely for the purpose of establishing a control group. However, if we are contemplating a potentially injurious manipulation, special care should be taken so that neither group, experimental or control, suffers because of threatening circumstances.

Often, evaluators are not allowed to establish control groups because it is said that the control group is at a disadvantage relative to the experimental group. Many state attorney generals would object to allowing potentially impaired drivers free access to the roads without undergoing medical review. Such a ruling, however, if carried to its logical extreme, would indicate that the medical review board did indeed improve the safety of medically impaired drivers and other motorists. The very ruling, in essence, obviates the evaluation process and defines the medical review board, for all intents and purposes, as effective. While such a ruling by an attorney general would no doubt be substantiated in a court of law, the ruling is antithetical to scientific process.

It is recognized that political considerations frequently must override scientific requirements. The medical review case cited above

may be such an example. Too often, however, control groups are not established in experimental designs and, instead, less powerful quasi-experimental designs are often substituted in order to "get around" this omission. The result is that a less desirable, less powerful evaluation is carried out.

IV. AN IDEALIZED MODEL FOR CARRYING OUT EFFECTIVENESS EVALUATIONS AND SEVERAL EXAMPLES OF WELL-DESIGNED AND WELL-CONDUCTED EVALUATIONS

Idealized Model for Effectiveness Evaluation.

How should evaluations be carried out? On the face of it, this is a very simple and straightforward question. Generally, it is assumed that all evaluators ask themselves this question when they initiate a project. It is also generally assumed that evaluators have a definite and conclusive answer to this question. In truth, many evaluators never ask themselves this question. Instead, they proceed without thought to collect project-related data in the vain hope that at some future date these data will miraculously provide the solution to their evaluation problem.

Lack of a clear knowledge of how to proceed to carry out an evaluation, perhaps more than anything else, will result in an inferior final product. Without adequate planning, the wrong experimental design may be chosen, the wrong data may be collected, the wrong statistical test of significance may be applied, and the wrong conclusion may be reached -- if any conclusion is reached at all. Without adequate planning, the conclusions of the evaluation will be found wanting.

After having said that adequate planning of evaluations is a common and serious fault in the field of highway safety, it must now be stated that there is no simple and quick remedy which can be prescribed to eliminate this problem in the future. There is no recipe or flow chart which can be provided for each and every evaluation in the field of highway safety. The problems inherent in the evaluation of guardrails and crash cushions are basically different from the problems which will arise in an attempt to evaluate an alcohol safety action program or a mandatory seat belt law. The field of highway safety is simply too broad, the various countermeasures which have been instrumented are simply too diverse to allow for a simplistic algorithm which dictates the path for each evaluation. Were such an algorithm available, the

tasks of evaluators would be much simpler, and the quality and uniformity of evaluations would be enhanced.

While it is true that no one can dictate in the abstract how all evaluations should be carried out, it is equally true that there are some basic steps or procedures which should be followed in all evaluations. By following these steps, the evaluator is not guaranteed valid conclusions, but by neglecting these steps he seems destined to arrive at less substantive and less valid conclusions than might otherwise be the case.

The four steps to be considered in conducting an evaluation are: (1) statement of goals, (2) design and measurement, (3) inference, and (4) conclusion and recommendation.

1. Statement of goals.

Without goals, evaluation is a waste of time, effort, and money. If specific, measurable goals cannot be established, evaluation will only be a cosmetic endeavor with predictable lack of success. All too often, attempts are made to measure the immeasurable -- the world of the empirically verifiable is abandoned; goals are set which are not measurable or even perceivable. This does not mean that only those programs which can be calibrated should be instituted. The unstated and the immeasurable can be pursued, but no attempts to evaluate such efforts for effectiveness should be made. Evaluation is an activity which demands certain prerequisites. Without goals (a dependent variable to measure), evaluation is both irrelevant and inappropriate.

An evaluator obviously cannot single-handedly determine the goals of a program. He should insist on clarity from the beginning on the part of administrators concerning their objectives. It is crucial that evaluation not be an afterthought. If evaluation is ignored in the planning stage, objectives will be vague, inflated, and tentative. Program people must be educated to the fact that measurement can only take place when there is some ever-fixed dependent variable to be measured. This does not mean rigidity, only that a program design is essential if it is ever to be determined that a program is working.

2. Design and measurement.

Once the dependent variable is defined, the next decision to be made is which experimental design should be used. The types of design which might be employed for an evaluation are numerous. Six designs have already been discussed. Unfortunately, no simple answer

is available to the question of which design is most appropriate. Often the practicalities of a given evaluation dictate which design must be used. But the findings which an evaluation yields are closely tied to the design which is chosen. Obviously, the choice of an appropriate design is of paramount importance.

3. Inference.

Most people see the third step of the evaluation model as the most complicated. Certainly, this step is the most mathematical; however, it is the most direct.

When a highway safety countermeasure is introduced, it is rare that all members of the treatment group fare better than all members of the control group. Instead, the real world dictates that some die in spite of the countermeasure, while others live without it. Some people wear seat belts and are killed, while others go beltless and survive severe crashes.

If all members of the treatment group were better off than all members of the control group, the third step of this model would be rendered unnecessary. Conclusions concerning the effectiveness of a countermeasure could be directly drawn. But such is not the case. Instead of dramatic differences between treatment and control groups, very small differences often occur. The group which received treatment X might have a fatality rate of 4.0 deaths per 100 million miles, while the control group has a death rate of 4.2 deaths per 100 million miles. Was the difference due to treatment or chance variation? Questions such as this lie in the domain of inferential statistics.

While different inferential statistics are available to be used with different types of data drawn under different experimental conditions, all inferential statistics attempt to show, with a given level of certainty, whether or not a particular difference should be attributed to chance or treatment. Inferential statistics is an inductive procedure whereby the relative odds of alternative explanations are pitted against each other. If a treatment group and a control group are seen to differ by a certain amount, and if through statistical procedures it can be shown that there was a realtively low probability that this difference would have occurred by chance, then it is inferred that the difference results from the treatment. In the language of inferential statistics, it would be said that a statistically significant difference existed.

Several points about statistical inference should be recognized:

1. A statistically significant result does not prove that the treatment was effective.

2. The fact that a statistically significant difference is not found does not necessarily mean the treatment is ineffective.

3. The fact that a statistically significant difference exists between the two groups does not necessarily mean that the program (treatment) is effective in any practical sense, i.e., "a difference is a difference only if it makes a difference" (Huff, 1954, p. 58).

4. Conclusion and recommendation.

> The results of the evaluation study to have any meaning at all must be translated into judgments of program success and failure (Suchman, 1967, p. 162).

Ideological differences between evaluators and administrators become apparent during the initial phases of evaluation and continue throughout the process. However, it is during the conclusion and recommendation portion that they become most critical.

> Right now a safety programme administrator is cast rather in the role of a football coach. He is not expected to experiment, he is expected to win! And if he doesn't, he's out. A safety programme is administered with sincerity and good faith, an effectiveness analysis is carried out, it is found not to be a sufficient impact and there is a tendency to blame the administrator. (Campbell, B.J., 1974).

It is asking a great deal of an administrator to heed the results of an evaluation which suggests that his program is of little or no value. In fact, it is asking a great deal of an administrator to allow his program to be scrutinized by an evaluator, knowing that such a conclusion might be reached. But if highway safety programs are to improve in the future, they must be evaluated today. Administrators must be convinced of the long run values of evaluation.

The evaluator, for his part, can foster this conviction by proceeding with tact throughout the evaluation period, and particularly during the conclusion phase.

Several rules should be followed by the evaluator in presenting his conclusions:

(a) The administrator should not be "surprised" by the conclusions contained in the written evaluation. The evaluator should be in close contact with the administrator throughout the study. If the data show that the program is worthless, the administrator should not be appraised of this in written report form. He should be prepared for this before the report is written.

(b) "...when the evaluator is drawing on knowledge and values outside the evaluation, he has a responsibility to say so. It behooves him to explicitly indicate the extent to which the recommendations he offers are supported by study data, how far they are logical extensions of the data, and where he has taken off on his own..." (Weiss, 1972, p. 126).

(c) Preferably, the conclusion should be written in clear, concise language. Evaluations which conclude with phrases such as -- 'the difference between the control group and the treatment group was shown to be significant at the .05 level' -- are indicative only of lack of literacy on the part of the evaluator.

The academic orientation sometimes leads evaluators to stop short of drawing conclusions when they report their results. As they see it, their job is to conduct the study and analyze the data; it is not to recommend action. Since the implications of data are rarely obvious, the evaluator's abdication of this task all too often means that nobody does it. The program manager winds up complaining about the irrelevance of the evaluation for his programmatic concerns, and the evaluator retires to his office lamenting the neglect of his work by decision makers (Weiss, 1972, p. 111-112).

Summary.

Evaluations should be seen as a synthesis between pure science and the politics of public administration. A creative combination of both will allow for clear statements of project goals, good measurements, logical inferences and subjective explanations for why things are the way they are and how they can be profitably changed. Objectivity in measurement, logic and common sense in inference, and subjective, qualitative prescriptions in conclusions are the essence of good evaluations.

Examples of Well-Designed and Well-Conducted Evaluations

1. Kaestner, N., Warmoth, E.J., and Syring, E.M. "Oregon Study of Advisory Letters: The Effectiveness of Warning Letters on Driver Improvement."

Purpose:

Advisory letters are sent by Departments of Motor Vehicles to notify individual drivers that their driving performance is under scrutiny and to encourage those drivers to improve their driving behavior. This latter function of the advisory letter is evaluated in the present study.

Method:

Some 944 male drivers who were eligible for advisory warning letters were randomly assigned to four groups: (1) control -- no letter, (2) standard form letter, (3) personalized standard letter, and (4) personalized, soft-sell letter. Members of the control group who were eligible to receive the warning letters were not notified that their driver record was being monitored by DMV. The control group was simply allowed one more than the usual number of traffic involvements before receiving a letter, and most of them did not receive warnings in the next 12 months. Group 2 received the standard letter. Group 3 received the standard letter, but the letter appeared to be individually typed and signed. The fourth group received a "personalized" letter that was more encouraging and less threatening than the standard letter.

Because it was not possible to avoid further contact with DMV during the study period, total traffic involvements could not be

directly compared. In other words, subsequent interviews, suspensions, and driver improvement school attendance could not be averted when further violations occurred. Therefore, all records surveyed were categorized as: (1) successes, (2) violation failures, or (3) accident failures at 6 and 12 month intervals. Success meant no entries, only minor violations, or nonchargeable accidents. Violation failures involved moving violations of a relatively serious nature. Accident failures were defined as chargeable avoidable accidents. The proportions of each group falling into these categories were compared.

Results:

The results indicated that those receiving the standard letter had the same subsequent driving records as the control group. But those receiving either the personalized standard letter or soft-sell letter had significantly more "successes" than the control group after 6 months. After 12 months, the soft-sell letter showed the most benefit, especially in terms of accident reduction. These results were accounted for primarily by drivers under 25 years of age who were no different from older drivers in previous violations and accidents but who responded more to the personalized letters by improved records. These results indicated the possibility of improving non-verbal (driving) behavior by verbal appeal in other ways such as interviews. (See Figures 6, 7, and 8).

2. Robertson, L.S., Kelley, A.B., O'Neill, B., Wixom, C.W., Eiswirth, R.S., and Haddon, W.,Jr. "A Controlled Study of the Effect of Television Messages on Safety Belt Use."

Purpose:

The present study sought to determine if the television medium could be used to increase seat belt wearing rates.

Method:

This study was carried out in a county with a population of approximately 230,000. Located within that county, there were some 13,800 households serviced by one of two cable television systems. Cable A serviced 6,400 households, and Cable B serviced 7,400 households. Demographically speaking, households on Cable A were indistinguishable from households on Cable B.

During the months June -- December 1971 and January -- February 1972, 943 safety belt messages were aired on Cable A, but not on Cable B. "If this campaign had been sponsored on a national basis,

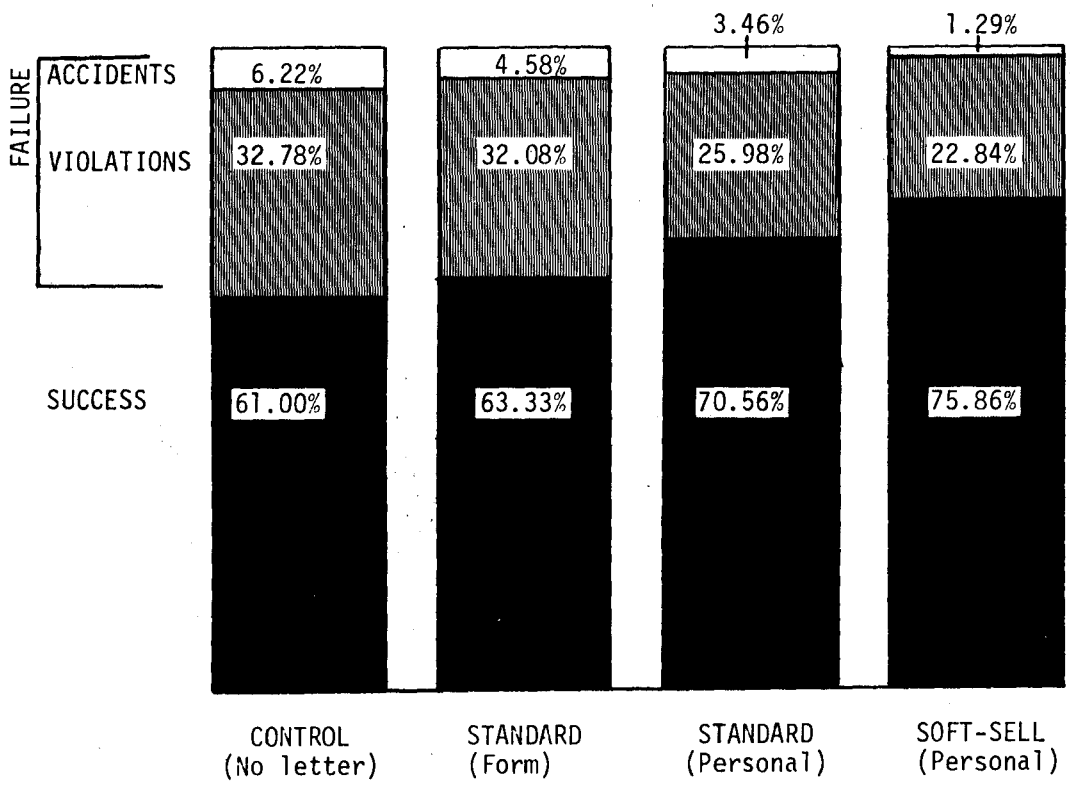


Figure 6. Six month comparisons of letters.

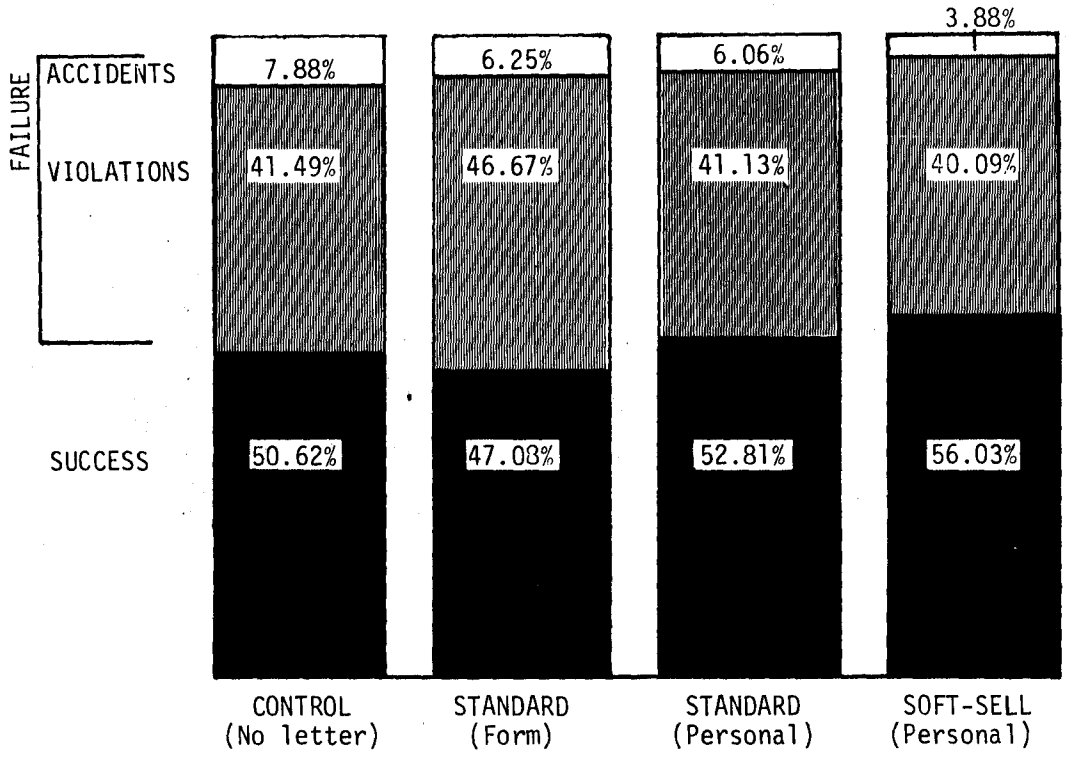


Figure 7. Full year comparisons of letters.

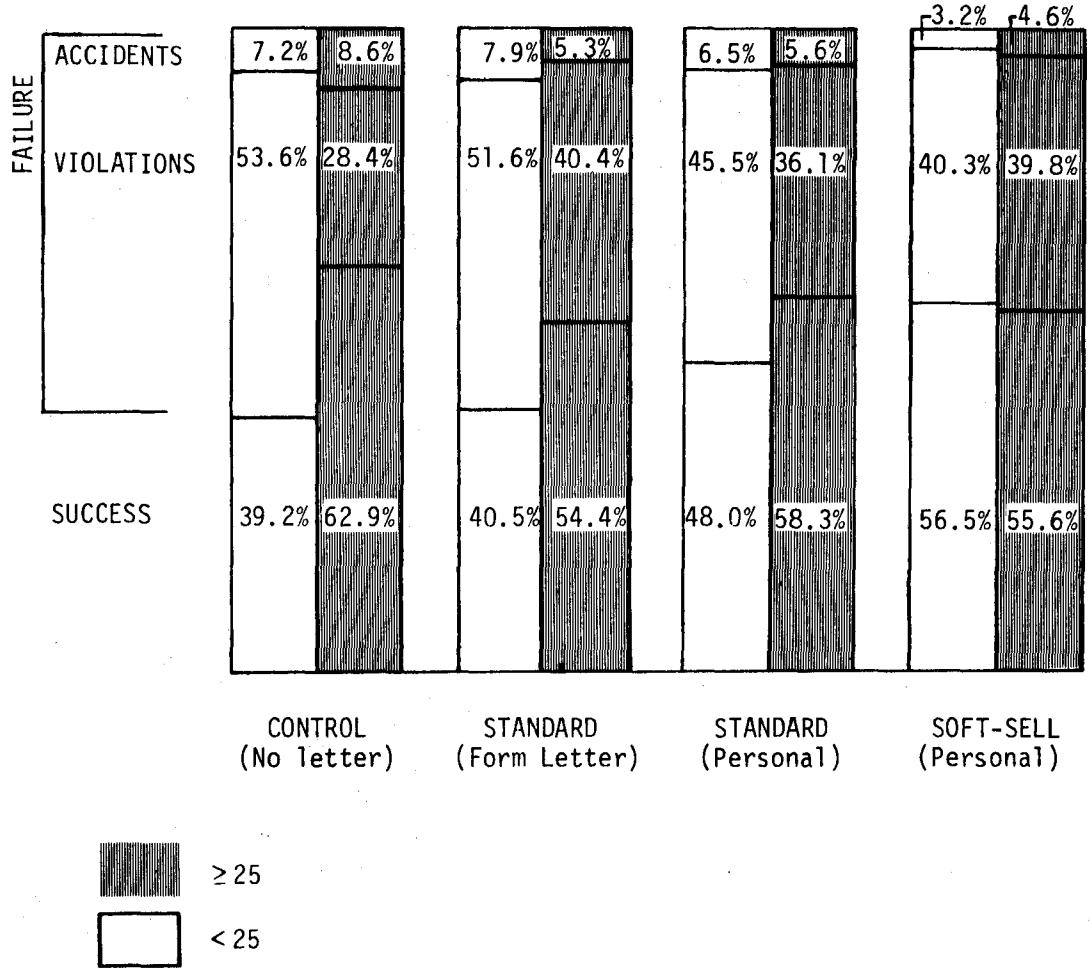


Figure 8. Comparison of letter effectiveness between younger drivers (under 25) and older drivers (25 and over) at one full year.

it would have cost approximately $7,000,000" (p. 9). Throughout those months and continuing through March, 1972, observers located at 14 different locations around the county recorded seat belt wearing rates for the drivers of passing motor vehicles. The observers themselves were rotated among the observation sites and, furthermore, they were unaware that their observations were related to any television campaign.

In addition to seat belt information, each observer simultaneously recorded the license plate number of the vehicle in question. By means of Department of Motor Vehicles files, individual vehicles were traced to specific street addresses. And on the basis of street address it could be determined that a particular vehicle belonged to a household serviced either by Cable A or Cable B, or neither.

The comparison of interest is, quite obviously, seat belt wearing rates for drivers from Cable A households compared to seat belt wearing rates for drivers from Cable B households.

Results:

The campaign had no effect whatsoever on safety belt use (p. 9).

3. Jones, M.H. "California Driver Training Evaluation Study."

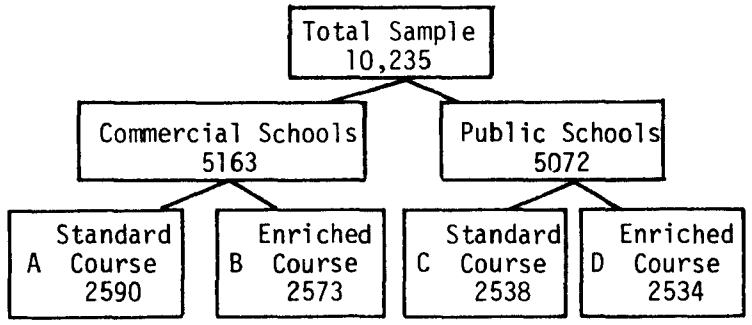
Purpose:

This study was carried out to determine if (1) different driver training methods (public school instruction versus commercial driving school instruction), and/or (2) enrichment of driver education programs could effect a reduction in accidents or violations in the 16-year-old driver. Enrichment was defined as four additional hours of instruction. The study was concerned only with the "skill acquisition" phase of instruction, the students having completed or having already been enrolled in the regular high school classroom driver instruction.

Method:

In order that the study results might be generalizable to the whole state, school districts to be used in the study were randomly chosen, to the maximal extent feasible. Similarly, commercial driving schools were randomly selected within the school districts of choice. And students were randomly assigned to the various experimental treatment groups.

In order to be 90 percent certain of detecting a two percent change in accident rate during a 12-month period (α set at 0.10 and assuming a standard deviation for accident rate no larger than 0.35), a sample size of at least 10,000 subjects was required. In all, some 10,235 subjects were used in the main analysis, as shown below.



Any difference in commercial school performance versus public school performance was determined by comparing accident and violation rates of A + B to accident and violation rates of C + D. The effect of the enrichment manipulation was determined by comparing accident and violation rates of A + C to the accident and violation rates of B + D.

Results:

> In sum, the conclusions of this study are clearer than had been anticipated. There are no differences in the essential criterion, accident rates, between public and commercial instruction or between standard and enriched programs (p. 16).
>
> (For further detail, see Table 1).

4. Andreassend, D.C. "The Effects of Compulsory Seat Belt Wearing Legislation in Victoria."

Purpose:

On December 22, 1970, legislation was passed in the state of Victoria, Australia, stating that "a person shall not be seated in a motor car that is in motion, in a seat for which a safety belt is provided, unless he is wearing the safety belt and it is properly

Table 1. Comparison of driving records by commercial vs. public driver training and by standard vs. enriched driver training (adapted from Tables 1.24A and 1.24C).

	Violation Rate		Accident Rate	
	Within 6 mos. of Licensure	Within 1 yr. of Licensure	Within 6 mos. of Licensure	Within 1 yr. of Licensure
Commercial	.173	.393	NS	NS
Public	.154*	.351**		
Standard	NS	NS	NS	NS
Enriched				

* $p < .10$, two tailed t test

** $p < .05$, two tailed t test

adjusted and securely fastened." Enforcement of this law began one month later. It was after this fact that the Road Safety and Traffic Authority (RSTA) was called upon to evaluate the new law.

Method:

In order to determine the effectiveness of the legislation, several procedures were undertaken. The main analysis, however, involved a time series design. Driver deaths which occurred in Victoria during the first six months of each calendar year (1955-1971) were plotted as a function of calendar year. Then a linear trend line was drawn through the first 16 data points.

After the trend function had been plotted for Victoria, the whole procedure was repeated for all other states in Australia exlucing Victoria.

Results:

By looking at the data shown in Figure 9 below, it was seen that for the several states in Australia excluding Victoria, the number of driving deaths is not appreciably below expectation, i.e., below the linear trend line. Driver deaths in Victoria, however, can be shown to be significantly less than expected for the first six months of 1971.

5. Campbell, B.J. "Seat Belts and Injury Reduction in 1967 North Carolina Automobile Accidents."

Purpose:

In 1964, seat belts became standard equipment in American cars. The question was asked: "How much protection does the seat belt afford the wearer?"

Method:

A total of 8713 crash-involved vehicles were sampled for purposes of this study. Drivers in 823 of the vehicles were wearing seat belts. Drivers in the remaining 7890 vehicles were not wearing seat belts.

Since the drivers of the crash-involved vehicles were not randomly assigned to the belt and non-belt groups, it could not be assumed that both groups were involved in similar accidents. And if,

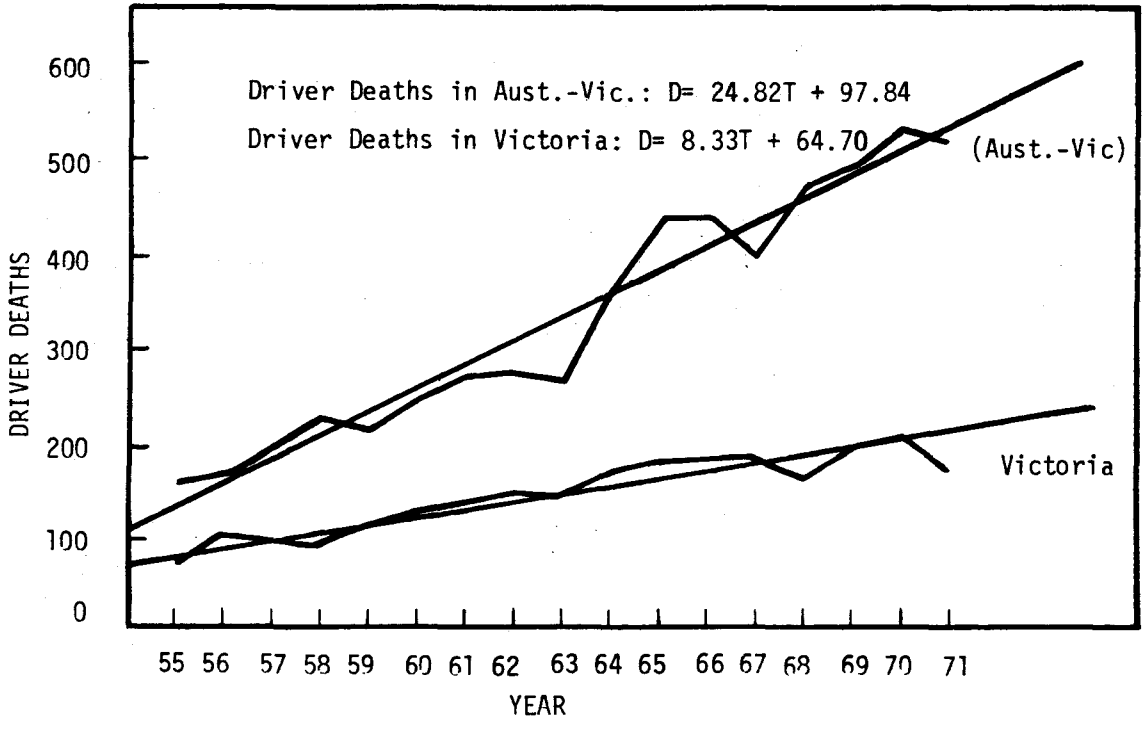


Figure 9. Trend line and actual driver deaths as a function of calendar year for Victoria versus rest of Australia.

for example, the belted group had been involved in relatively more benign accidents, any statement regarding seat belt benefit would have been inflated.

In order to get around this potential source of bias, each of the 8713 crash-involved vehicles was categorized according to accident circumstances. Some vehicles were found to be travelling at slow speeds when they ran off the road and struck a fixed object with the front end of the vehicle. Other vehicles were found to be involved in head-on collisions with trucks while travelling at high rates of speed. All 8713 vehicles were categorized into one of 140 different classifications of accident circumstances. Each one of these 140 different classifications was associated with a certain probability of death or serious injury for the unbelted driver and a different, usually lower, probability of death or serious injury for the belted driver.

In order to determine overall seat belt effectiveness, the probability of serious injury or death for belted drivers in each of the 140 different categories was weighted (multipled) by the relative frequency of occurrence of the 140 categories among the non-belted drivers. Then these weighted probabilities of serious and fatal injuries for belted drivers were summed. The resulting overall probability of serious injury or death for belted drivers was now unbiased (at least in terms of the 140 categories) with respect to the probability of death or serious injury of unbelted drivers. At this point, it was a simple matter to determine seat belt effectiveness by subtracting probability of serious injury or death for belted drivers from serious injury or death for unbelted drivers and then dividing that difference by the probability of serious injury or death for unbelted drivers.

Results:

The results of the steps described in the Method Section indicated that seat belts reduced the chance of serious or fatal injury by 36 percent.

V. CONCLUSIONS AND RECOMMENDATIONS

The underlying theme of this paper has been that many impediments exist to the evaluation of highway safety programs. It was pointed out in Chapter I that the state of the art of highway safety evaluation could be advanced only if the many impediments to evaluation could be combined into categories and addressed in groups. Chapters II and III concentrated on detailing the external and internal impediments to the evaluation process itself. Chapter IV sought to provide examples of evaluations which were well executed.

On the basis of the categorization which was established in Chapter I, several recommendations can be made which, if carried out, would improve the state of the art of highway safety program evaluation.

Recommendations for Overcoming Impediments to Evaluation

Naive ignorance.

Chapter II pointed out that many administrators, legislators, and private citizens are benignly ignorant of the evaluation process itself. Other administrators, legislators, and citizens have a passing knowledge of the rudiments of evaluation but are unconvinced of the power of the process in terms of savings in lives and dollars.

In order to overcome the impediment of naive ignorance on the part of these various groups, it seems clear that the main remedial theme should be education. Administrators, legislators, and citizens who are benignly ignorant of evaluation might well adopt its principles if they were made aware of its existence.

In order to educate these groups, several avenues are open. It is anticipated, however, that the largest group (individual citizens) will be the most difficult to reach. Legislators, due to their reduced number, will be somewhat more approachable. Finally, administrators, having the fewest numbers, should be the most accessible.

Specific recommendations.

1. Most secondary school systems in the United States contain a course in American government entitled Civics, Problems of American Democracy, etc. Many secondary school systems offer courses in economics and sociology. It seems reasonable to assume that the formats of any of these courses could be expanded to include a block of instruction on evaluation principles in general. Such a block of instruction on evaluation could take examples from fields such as highway safety, education, public policy, etc.

In order to foster the teaching of evaluation in the high schools, it is felt that a basic text (perhaps only a chapter within a text) could be undertaken by various publishers or professional organizations. The intent of this text would not be to produce accomplished evaluators at the high school level, but instead to produce more educated, better informed, more questioning voters.

2. In order to reach those citizens who are already past secondary educational levels, presentations to various civics groups and pressure groups would seem in order. Again, the intent of these presentations would not be to produce accomplished evaluators, but instead to familiarize large segments of society with the evaluation process.

3. Legislators, the point of monetary origin for nearly all highway safety projects, are ignorant of evaluation procedure. They fund programs which should not be funded, and they pass over programs which might be beneficial.

Many states provide an orientation program for freshmen senators and representatives. This orientation is frequently carried out at some state institute of government or within the state's university system. Such an orientation would provide ample opportunity for basic instruction in the tenets of program evaluation.

4. The first chapter indicated that the states are not adequately evaluating programs which are funded under Section 402 of U.S. Code, Title 23. In order to overcome this lack of evaluation at the state level, Governor's Highway Safety Representatives and their staffs need to be grounded in fundamental program evaluation procedures. Workshops for this group should be considered by the federal government, various trade organizations, and the National Association of Governor's Highway Safety Representatives itself.

Again, the format of these workshops should be set up to train people to appreciate the evaluation process and not necessarily to become professional evaluators. If it were felt that a given staff member within a Governor's Highway Safety Representative's office would serve as the professional evaluator of projects, then he should take more extensive evaluation training in more advanced workshops or within a university curriculum.

Additionally, it would seem incumbent on the regional offices of the National Highway Traffic Safety Administration to provide information to the states regarding why evaluations are important, what constitutes a good evaluation, how evaluations should be carried out, etc.

5. In keeping with the need to educate state administrators and their staffs, a handbook on program evaluation would be most useful. The format of such a handbook should address itself to the basic principles underlying evaluation; it should explain in clear, non-mathematical terms how an evaluation is carried out; it should explain the usefulness and benefits to be derived from evaluations. The handbook should not concern itself with such things as: on what government form should the evaluation be written, how is Chi-square calculated, or what is the most efficient means of drawing a stratified random sample. The handbook should be more philosophical in flavor. It should not be a treatise on statistics or a manual on how to fill out government forms. A handbook such as the one just described might best be written through funds appropriated by the National Association of Governor's Highway Safety Representatives or various trade organizations.

Administrator wisdom.

While many program administrators and governor's highway representatives are ignorant of the evaluation process, many others, perhaps a majority, are aware of program evaluation and are actively opposed to it. These administrators are aware that evaluation is a time-consuming process which necessarily impinges upon their staff time, which interferes with staff procedures, and which often results in a negative finding. Under these circumstances, the reluctance of an administrator to carry out an evaluation is quite understandable.

If administrators are to be prevailed upon to produce good, rigorous evaluations, a new set of contingencies will have to be established for these administrators. While simple educational measures may be effective with the naively ignorant administrator, they

will be totally ineffective with any administrator who is actively resistant to the evaluation process.

In order to overcome the impediment of administrator resistance to evaluation,the basic theme which should be employed is sanctions. At the present time, the federal government dispenses money to the states on a federal-state matching basis in order that they might fund projects which tend to bring states in line with the standards promulgated by NHTSA. Each program which is funded under these provisions carries with it the obligation for evaluation. But, as was pointed out in Chapter 1, this obligation is rarely met or met very inadequately. NHTSA is empowered to withhold funds (402 funds and 10 percent of federal highway funds) from any state which is not actively moving toward compliance with the program standards. If this power is interpreted in a broad sense, then essentially every state in the nation is non-compliant and from this it follows that funds could be withheld.

Such a move on the part of NHTSA would be politically inexpedient. It is not suggested that this move should be taken, at least on a broad scale.

Specific recommendations.

1. If NHTSA is to be in a politically expedient position to exert power on the states to produce good program evaluations, it should endeavor, first of all, to put its own house in order.

   a. All of the program standards promulgated by NHTSA carry the requirement that programs under the standards shall be evaluated. Unfortunately, the word evaluation is poorly defined within the standards and great leeway is thus allowed individual states in arriving at their own definition. NHTSA should endeavor to define in very explicit terms what they will accept as an adequate evaluation from the states.

   b. Some of the standards which require evaluation are totally unamenable to Type III (effectiveness)evaluation. Traffic records programs, for example, which are required of the states, are not designed to directly reduce accidents, injuries, fatalities or any other measurable dependent variable. Accordingly, programs funded under this standard should be evaluated via the tenets of the Type II evaluation. When and if NHTSA defines what it means by evaluation, they should be clear to point out this distinction between a Type II and Type III evaluation, and

they should further make clear which evaluation process should be used with which programs.

c. Many of the state administrators who actively resist evaluation do so not because a negative evaluation will result, but in the firm knowledge that the results of any evaluations which they submit will not be acted upon. The standards themselves, for example, have become "ever-fixed marks" which have become sacrosanct and seemingly unchangeable. How many programs funded in the name of a given standard must show negative results before someone questions the validity of the standards? This is a question which state administrators ask themselves. Realizing that evaluations rarely change the course of program implementation, state administrators feel that evaluation is nothing other than a paper process which is of little or no consequence.

Until such time as NHTSA is willing to amend its standards, or abolish some of them altogether, this feeling of frustration on the part of state administrators seems likely to continue. And if NHTSA does not act responsibly on the evaluations which they receive, then indeed the state administrators correctly perceive the whole process as a bureaucratic waste of time and money.

d. Not all programs should be evaluated, at least via the Type III procedure. If a program has been shown to be effective in many cases, further evaluations seem unjustified. Similarly, if a given type of program has been shown to be worthless on many different occasions, programs should not be instrumented and evaluated.

2. Assuming that NHTSA carried out the Steps 1.a-d, they would be in a stronger, more tenable position in insisting on more rigorous evaluations from the states. Once they have defined what they will accept in the name of evaluation, they should rigorously insist on compliance. If, for example, a given state funded a particular organization to carry out a program and if that organization inadequately evaluated the effects of their program, future funding for that organization should be suspended. Note that the organization would not be suspended for a negative evaluation, but for an inadequate evaluation.

All organizations applying for grants under 402 funds should be required to specify in very specific terms how they intend to go about evaluating their program. Glittering generalities such as "the program will be evaluated by experts," or "we will put the data up on computers," etc. are of little consequence. The

evaluation plan should contain such things as: What are the program goals? How are the goals measured? How will the control group be constituted? Has the evaluator made provisions for blind or double-blind procedures, if necessary? Questions such as these are of consequence, and they must be stated a priori, before the program is put into effect. Any proposed program lacking a clear and methodologically sound evaluation plan should not be funded.

3. Programs which are of considerable size, e.g., over $100,000 for a given fiscal year, should be contracted out to organizations with some expertise in the field of evaluation. While evaluation itself is basically a simple and straightforward process, it is, nevertheless, a simple and straightforward process which relatively few organizations have mastered. To allow an organization without the requisite experience to perform evaluations may be wasteful in the long run. Furthermore, by having an organization evaluate a program from the outside, less biased, more valid results may be achievable.

Technical ignorance.

At the present time, researchers and evaluators in the field of highway safety are drawn from many disciplines -- engineering, psychology, statistics, mathematics, operations research, etc. There is no master's degree or Ph.D. degree in highway safety or highway safety evaluation. Rather, the people who come into this field learn certain experimental, epidemiological, and statistical methods which are appropriate to a wide variety of subject matters. Unfortunately, this field has not attracted the quantity or quality of professionals which would be desirable. Manpower in highway safety and highway safety evaluation is seriously lacking.

The main theme to redress this impediment is quite obviously education.

Specific recommendations.

1. NHTSA should fund various colleges and universities to undertake to produce competent evaluators at the undergraduate and graduate levels. Evaluation curricula could be established within departments of applied psychology, public health, public administration, etc. These curricula could be enriched if they were tied in with existing highway safety centers at uniersities around the country (e.g., Texas A & M, University of Indiana, the University of Michigan, the University of North Carolina, University of Southern California, etc.).

2. In addition to funding specific colleges and universities, NHTSA should offer scholarships or assistantships to qualified individuals in highway safety and highway safety evaluation. These individuals could be chosen on the basis of merit, and they could be granted funds to carry out a curriculum agreeable to the student, the university, and NHTSA.

3. In order to foster highway safety program evaluation and in order to lend a flavor of credibility to the field, it would be desirable to have an endowed chair of highway safety or highway safety evaluation established by some philanthropic organization or trade organization. Such an endowed chair would tend to draw students at the graduate level into this field, and it would allow the individual holding the chair to express his own views, beliefs, and findings without threat of repercussions from industry or government.

4. While it is desirable to have more and better trained professionals brought into this field, it is recognized that numerous professionals and paraprofessionals are presently acting to carry out evaluations at this time. To reach these individuals, it would seem appropriate that a series of technical, evaluation workshops be put together. Participants in these workshops would come from within NHTSA and from the individual states. The participants would be expected to have some prerequisite abilities in the field of highway safety evaluation, including such things as basic familiarity with subject matter in the field, at least one course in statistics or experimental design, etc.

5. In order to carry out a workshop described in the immediately preceding paragraph, it would be most desirable to have a textbook on the science of highway safety evaluation. This book would no doubt have chapters on such subjects as statistics, sampling procedures, experimental design,etc. The text should be written at a reasonably technical level and would be aimed at an audience of senior level undergraduates or first-year graduate students.

Such a text would be beneficial for purposes of training evaluators already in this field, and it would be of benefit in training individuals within universities who have decided to enter or who are considering entering the highway safety/highway safety evaluation field. Ideally, it would be hoped that such a text might be commercially viable and thus underwritten by a publisher. If such were not the case, however, it is hoped that such a text

could be underwritten by the federal government, a trade organization, or a university press.

6. Finally, it would be hoped that the field of highway safety evaluation has advanced to such an extent that it could support a journal of its own. At present, there are several safety journals in existence, and there is at least one journal in existence which specifically pertains to evaluation. Perhaps the time has come that a journal of highway safety evaluation is appropriate. Such a journal would ideally contain theoretical articles, review articles, and articles pertaining to specific evaluations. By means of such a journal, professionals in the field would not only keep abreast of evaluation methods and techniques, but they would also learn what other evaluators in the field were doing and how they (i.e., other evaluators) were dealing with particular problems and difficulties.

It is suggested that this journal might be most effective if it were established within a university setting and funded through NHTSA, trade organizations, or some combination thereof. By its location within a university, it would be hoped that an editorial board could be established with requisite freedom and rigor to maintain a sound and viable journal.

Inadequate tools, procedures, data bases.

Many states at the present time are not equipped to carry out highway safety evaluations. Many states have traffic records systems which are so antiquated that it would be almost impossible to evaluate the effects of certain laws, enforcement procedures, etc.

Many states have not had a history of conducting highway safety evaluations, and therefore, they have not enlisted the aid of a highway patrol, local judges, or hospital personnel for purposes of carrying out evaluations. Many states are not allowed, at the present time, to establish necessary control groups within their population for purposes of evaluating a particular highway safety countermeasure.

All of these problems of data collection and data handling make the job of the evaluator more difficult. It should be noted, however, that if more people know about evaluation, if more people are more technically trained to carry out evaluations, and if the federal government insists upon evaluations, these impediments to the process of carrying out evaluations will disappear over a period of time. These

problems of inadequate data, inadequate control groups, etc. will be self-correcting. Self-correction will not occur overnight, but it will occur slowly and almost necessarily, if the impediments of naive ignorance, administrator wisdom, and technical ignorance can be overcome.

REFERENCES

American Psychological Association, Inc. Ethical principles in the conduct of research with human participants. Washington, D.C.: Author, 1973.

Andreassend, D.C. The effects of compulsory seat belt wearing legislation in Victoria. Paper presented at the National Road Safety Symposium, Canberra, March 1972.

Brooks, M.P. The community action program as a setting for applied research. In F.G. Caro (Ed.), Readings in evaluation research. New York: Russell Sage Foundation, 1971.

Campbell, B.J. Seat belts and injury reduction in 1967 North Carolina automobile accidents. Chapel Hill: University of North Carolina Highway Safety Research Center, 1968.

Campbell, B.J. Highway safety program evaluation and research. Traffic Digest and Review, 1970, 18 (1), 6-11.

Campbell, B.J. Driver injury in automobile accidents involving certain car models. Chapel Hill: The University of North Carolina Highway Safety Research Center, July 1970.

Campbell, B.J. Highway safety progress during recent years. In Proceedings: Automotive Safety Engineering Seminar, June 20-21, 1973, Warren, Michigan.

Campbell, B.J. Driver injury in automobile accidents involving certain car models: An update. Chapel Hill: University of North Carolina Highway Safety Research Center, 1974.

Campbell, D.T. Measuring the effects of social innovations by means of time series. In J.M. Tanur, F. Mosteller, W.H. Kruskal, R.F. Link, R.S. Pieters, G.R. Rising (Eds.), Statistics: A guide to the unknown. San Francisco: Holden-Day, Inc., 1972.

Davis, H. A solution for crisis? Evaluation, 1972, 1 (1), 3-5.

Governor's Highway Safety Program. Highway safety program standards. Raleigh: Author.

Greenshields, B. Traffic and highway safety and how it may be improved. Science, 1970, 168, 674-78.

Griffin, L.I. The effects of population patterns on the motor vehicle death rate. Highway Safety Highlights, June 1974, 8 (2), 2-3.

Guilford, J.P. Psychometric methods. New York: McGraw-Hill Book Company, Inc., 1936.

Herms, B.F. Pedestrian crosswalk study: Accidents in painted and unpainted crosswalks. San Diego, California: Public Works Department, Traffic Engineering Section, 1970.

Huff, D. How to lie with statistics. New York: W. W. Norton & Company, Inc., 1954.

Jacobs, H.H. Conceptual and methodological problems in accident research. In H. H. Jacobs, E. Suchman, et al. Behavioral approaches to accident research. New York: Association for the Aid of Crippled Children, 1961.

Joksch, H.C. A comprehensive search for cost-effectiveness data for highway safety countermeasures. Hartford, Connecticut: The Center for the Environment and Man, Inc., 1972.

Jones, M.H. California driver training evaluation study. Final report. Los Angeles: UCLA School of Engineering and Applied Science, 1973.

Kaestner, N.F. The impact of driver improvement: Do we really want to know? In P. F. Waller (Ed.), North Carolina Symposium on Highway Safety. Vol. 10. Highway safety programs: How do we know they work? Chapel Hill: University of North Carolina Highway Safety Research Center, 1974.

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.

Kelvin, W.T. Popular lectures and addresses, by Sir William Thomson. New York: MacMillan and Co., 1894.

Kershaw, D.N. A negative income tax experiment. Scientific American, 1972, 227 (4), 19-25.

Malfetti, J.L., & Simon, K.J. Evaluation of a program to rehabilitate drunken drivers. Traffic Quarterly, 1974, 28, 49-59.

National Safety Council. Accident facts: 1974 edition. Chicago: Author. 1974.

New York Times, March 10, 1974.

Reese, J.H. The legal nature of a driver's license. Washington,D.C.: Automotive Safety Foundation, 1965.

Robertson, L.S., Kelly, A.B., O'Neill, B., Wixom, C.W., Eiswirth, R.S., & Haddon, W.,Jr. A controlled study of the effect of television messages on safety belt use. Washington, D.C. Insurance Institute for Highway Safety, 1972.

Ross, H.E., & Post, E.R. Criteria for guardrail need and location on embankments. Volume I: Development of criteria. College Station, Texas, Texas A & M University, 1972.

Ross, H.L. Interrupted time-series methods for the evaluation of traffic law reforms. Paper presented at North Carolina Symposium on Highway Safety, Spring, 1974.

Salasin, S. Experimentation revisited: A conversation with Donald T. Campbell. Evaluation, 1973 1 (3), 7-13.

Sidman, M. Tactics of scientific research, evaluating experimental data in psychology. New York: Basic Books, 1960.

Stoke, C.B. Reflectorized license plates: Do they reduce nighttime rear-end collisions? Charlottesville: Virginia Highway Research Council, 1974.

Suchman, E.A. Evaluative research. New York: Russell Sage Foundation, 1967.

Swinehart, J. & Grimm, A. (Eds.) Public information programs on alcohol and highway safety. Ann Arbor: University of Michigan Highway Safety Research Institute, 1972.

Tourin, B., & Garrett, J.W. Safety belt effectiveness in rural California automobile accidents: A comparison of injuries to users and non-users of safety belts. Buffalo: Automobile Crash Injury Research of Cornell University, February 1960.

Weiss, C.H. Evaluation research. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1972.

Weiss, R.S., & Rein, M. The evaluation of broad-aim programs: A cautionary case and a moral. Annals of the American Academy of Political and Social Sciences, 1969, pp. 133-42.