The Effectiveness Evaluation of Transportation Projects: Three Case Studies in Pennsylvania

PHILIPPOS J. LOUKISSAS and JOHN L. MACE

ABSTRACT

Issues relevant to the conduct of effectiveness or postimplementation evaluation studies are addressed within the context of urban transportation projects. In particular, the conceptual rationale underlying development of the experimental plan is discussed in detail. An experimental plan serves to rule out systematically factors that might cause an observed difference between measures of effectiveness. Several such recurrent factors--called threats to internal validity--are described as well as are experimental plans useful in ruling them out. These more theoretical notions are discussed subsequently within the contexts of three transportation projects in Pennsylvania. These include a safety improvement project, a ridesharing program, and a transit fare reduction study. The findings of the study, which resulted from a review of the literature and interviews with transportation officials, suggest an unnecessarily wide gap between the state of the art and the actual practice in the design of effectiveness evaluations. The authors offer some suggestions for improving the conduct of such studies as a means of improving the way decisions are made, particularly at the state level.

Effectiveness evaluations are addressed as they relate to transportation projects and programs. An effectiveness evaluation is a postimplementation study to assess whether or not a particular project was successful in achieving its objectives. Effectiveness evaluations are contrasted with the preimplementation evaluation of alternative plans and with administrative evaluations of processes by which decisions are made and implemented. Within the transportation field, effectiveness evaluation is the least well understood and the least frequently practiced type of evaluation. Although there are many financial, institutional, and political explanations for this, insufficient technical training also inhibits the conduct of strong effectiveness evaluations. In particular, the design of such evaluation studies is not a part of the usual training received by the transportation engineers and planners who are responsible for conducting them. In the past few years, several documents in the transportation literature have addressed the technical issues underlying the development of evaluation plans (for example, 1-3). In the judgment of the present authors, these documents would be improved by a more precise explanation of what constitutes a strong effectiveness evaluation. In response, the purpose of this paper is to present the conceptual rationale underlying evaluation plans.

The information presented here has been gathered in two ways. The first was by reviewing relevant literature on evaluation methodology in the fields of transportation, planning psychology, and education. In addition, about thirty transportation professionals, representing the federal government, the Pennsylvania Department of Transportation, local transportation agencies in Pennsylvania, and private consulting firms were interviewed in person or by telephone.

There are three major sections in this paper. The first section is theoretical and presents the nature and structure of experimental plans or research designs. The theory is then applied through discussions of the evaluation plans associated with three transportation projects in Pennsylvania. The final section addresses other technical and nontechnical issues related to evaluation studies and discusses means for improving state level evaluation capabilities.

EXPERIMENTAL PLANS

An experimental plan, also called a research design, is the central element of an effectiveness evaluation. It is also the least well understood element. In psychology and education experimental plans are understood as arrangements of treatments and observations in time. A treatment is some intervention: a particular program or a set of countermeasures. An observation is the value of a measure of effectiveness (MOE).

One way to view the decision to implement a transportation project is to consider it as an attempt to change the value (single value, average value, or distribution) of an MOE by implementing a program or a set of countermeasures. Within this context, the purpose of an effectiveness evaluation is to answer three related questions.

1. Was the value of the MOE changed?
2. If so, did the program or countermeasures cause the change?
3. If the program caused the change, are the results applicable to other situations?

The first question is a statistical one. A statistical test is performed to determine if there is a significant difference between two or more values of the MOE. The second two questions, which are answered through the experimental plan, are the focus of this paper. Such a research design permits the evaluator to conclude that certain obvious factors besides the program or countermeasures did not cause the difference between MOEs. The more factors that can be eliminated by an experimental plan, the stronger or more internally valid that plan is considered to be. In attempting to answer the third question, the evaluator addresses the issue of external validity.

The True and the Quasi-Experiment

Current theory about experimental plans constitutes

a generalization of the rather refined methods used by laboratory scientists, who make explicit use of the technique of experimental control. Two primary ways in which experimental control is realized are random selection and random assignment. In a field situation it is difficult to conduct true experiments because the use of random selection and assignment may not be appropriate or feasible. In this case the researcher must systematically eliminate the possibility that other factors were responsible for the change by developing a quasi-experimental plan (4).

By strengthening the external validity of an evaluation plan, an evaluator is attempting to approximate the effects of random selection. Conversely, by strengthening the internal validity of a plan, the effects of random assignment are approximated. These two types of validity are discussed in more detail in the next section.

Internal Validity and Experimental Plans

In a nonexperimental situation, there are always numerous possible alternative explanations for observed differences between MOEs. Psychologists and educators working with D.T. Campbell (4) have classified types of such explanations that arise frequently. Six of these types are described below, along with ideas for experimental plans by which they might be ruled out. They are often referred to as threats to internal validity.

1. History. It is quite possible that some event other than the treatment under investigation resulted in all or part of any difference between observations at the project site. Consider, for example, a safety improvement project implemented during the same time period that the national speed limit had been reduced from 65 to 55 mph. If a reduction in the number of accidents was observed, an evaluator could have no way of indicating which event--the project or the speed-limit change--might have reduced the number of accidents. Generally, a comparable group or site likely to experience the same set of historical events between observation times is sufficient to "control" for history.

2. Maturation. Maturation is a natural change over time in the MOE being observed. Failure to account for a trend of accidents occurring at a site before a project is implemented makes it impossible to separate the effect of the project from the maturation effects. Controlling for maturation necessarily depends on detecting a preexisting trend or cycle.

3. Selection. As a threat to internal validity, selection is a problem whenever project groups and comparison groups differ in a significant way. If project sites are not selected for funding at random, there may be reason to suspect that they are different from the control group. In many situations, it is quite plausible that this difference could explain differences observed between treatment and control groups. To control for selection bias, it is important to match comparison sites to experimental sites and to estimate the impact of possible differences between the sites on the appropriate MOEs.

4. Interactions with selection. If the experimental group is different somehow from the comparison group, it may react differently to historical events or it may change differently over time. In other words, selection may interact with either history or maturation. As an example of a selection and maturation interaction, consider a high-accident project site and a comparison site, both having the same number of accidents during the year previous to project implementation. If the project was effective, one would expect the number of accidents at the project site during the year following implementation to be lower than the number at the control site during the same year. Suppose the number of accidents at the project site was increasing over time while the number at the control site was decreasing. In this case even if the accident rate at the project site were reduced, a faulty inference might be made that the project was ineffective. In this situation multiple observations at both sites before implementation of the project would help reduce the plausibility of this interaction effect.

5. Regression. Regression is the observed tendency of extreme scores to regress or move toward the average score. If a group of sites is selected as being eligible for funding because each has a high number of accidents, the average number of accidents at sites within this group will probably be lower the next time accidents are counted. The only way to account for regression effect is to estimate it by measuring the average decrease among all untreated members of the group.

6. Instrumentation. The instruments here are devices or procedures used to generate the MOEs. Instrumentation is a problem that arises when an instrument changes over time making comparisons of MOEs impossible. If an instrumentation problem is to be overcome, one must make some adjustments in the values of the observations to account for the changes in the instrument.

Although this list of common threats to internal validity could be made longer, major threats to transportation projects and programs have been covered. Familiarity with this list will allow the evaluator to determine those specific threats that are plausible for a particular evaluation. When the relevant threats have been determined, an appropriate experimental plan may be developed.

External Validity

In the absence of random selection, there are always threats to external validity. Generally speaking, these threats mean that there is something unique about the project situation that prevents the results from being applicable to other situations. Consequently, the results of the study might be due to an interaction between the project and the unique attributes of the experimental group or site (selection, setting, or history).

An additional threat to external validity occurs when the treatment itself changes over time. Many programs undergo revisions on a periodic basis. When the effects of a program are cumulative over time, it may not be clear what is being generalized. In such instances it is useful to separate the effects as closely as possible and to keep careful records of how the program changes. In this way, not only will effects be more easily associated with particular program or project elements, but the generalization of effects will be more valid. A related and prevalent difficulty arises when several treatments are applied at the same time. Basically, the difficulties with generalization may be overcome only by replication of the treatment to several representative groups at different sites and at different times.

REVIEW OF CASE STUDIES

A review of the literature and interviews with

transportation professionals at various levels of government revealed that, in general, there is no demand for evaluation studies beyond the administrative or clinical type (professional opinions of the efficiency with which bureaucracies perform their functions). In the transportation field, more formal evaluation studies are usually conducted in response to federal or state requirements, either as part of a strategy to justify a continuation of funding or to assess the effectiveness of federally funded demonstration programs.

Three transportation projects in Pennsylvania have been selected to illustrate experimental plans typically used for the purposes described previously. The first is a highway safety improvement project in State College anticipated by the district office of the Pennsylvania Department of Transportation (PennDOT), the second is a ridesharing program undertaken by the Delaware Valley Regional Planning Commission, and the third is a current transit fare reduction demonstration project in Scranton, Pennsylvania, conducted by the local transit authority. The responsibility for evaluation in both the safety improvement and ridesharing projects lies with the implementing agency. The transit fare reduction project, on the other hand, is being evaluated by the Transportation Systems Center. In two of the three cases, the evaluation studies had not been completed when this paper was written. In the first case, PennDOT indicated that the evaluation plan would be the type traditionally used in such projects. In the third case, the document describing the evaluation plan to be used was also evaluated.

These three cases represent three classes of transportation decisions: a physical modification, the institution of a program, and a policy change. Not surprisingly, the nature of the intervention and the study objectives have implications for the type of evaluation design that is appropriate. The purpose of discussing these cases is not to criticize particular evaluation designs. Instead it is to make more clear the ideas of internal and external validity by discussing them within familiar contexts and to assess the state-of-the-practice of evaluation studies.

A Safety Improvement Project: North Atherton Street, State College, Pennsylvania

PennDOT has funded a safety improvement project along a section of North Atherton Street in State College. This project is illustrative of a large class of physical countermeasure projects regularly undertaken by transportation agencies.

Interest in the North Atherton Street project began when PennDOT obtained a low skid reading that obligated the department to take some corrective resurfacing action. Because the 1.2-mile stretch of four-lane state highway has experienced a relatively high accident count over the past 5 years (420 reported accidents with no fatalities), it was considered for further funding as a safety improvement project.

In addition to resurfacing the entire project section, proposed design solutions included a 2-to-4 foot widening over most of the section to allow for a fifth left turn lane at several of the intersections. The resurfacing and channelization are expected to reduce the number of accidents appreciably. Because of minimal widening and the addition of a lane, the other lanes had to be narrowed from 11 to 10 feet. This narrowing is expected to reduce automobile speeds, thereby responding to a major concern of local residents.

According to the guidelines specified in FHPM 8-2-3 (5), when safety improvement projects are implemented an effectiveness evaluation is required. The evaluation content must have three elements: a statement of the project costs and safety benefits; a record of the accidents before and after project implementation; and a comparison of accident numbers, rates, and severities after implementation with those expected if the project had not been implemented.

In practice, at least in the Commonwealth of Pennsylvania, it is generally assumed that if the project had not been implemented, the before observations would remain unchanged. This evaluation plan is called a before and after design. To comply with the reporting procedures outlined in FHPM 8-2-3, the raw before and after data are summed over projects assigned to the same safety classification code and, as appropriate, summed over 1, 2, or 3 years. In its 1980 annual report (6) PennDOT does not draw any inferences about the effectiveness of the projects.

The before and after design is not internally valid even if it is assumed that a significant difference between MOEs (here number of accidents) has been observed. There are four serious threats to internal validity--instrumentation, maturation, regression, and history--that prevent any conclusion as to whether or not the project was responsible for the difference. For the moment, the threats of selection and interactions with it may be ruled out because only one group is specified in the experimental plan.

Additionally, there was a change in the accident reporting procedures. Under the new motor vehicle code enacted in 1977, a dollar threshold criterion for reportable property damage only (PDO) accidents was replaced by a towaway criterion. An accident that was reported in 1976 might not have been reported in 1978; consequently, the annual figures are not comparable. PennDOT has responded to this instrumentation bias by including only injury and fatal accidents in its before and after evaluations. The authors conducted a study for PennDOT in 1982 where procedures are recommended to make pre-1977 data comparable to post-1977 statistics (7).

The most serious threat to the internal validity of the before and after study is maturation. If the accident counts are increasing over time, then it is quite reasonable that the proposed project could be effective even if the after observation is higher than the before observation. An extended number of observations is necessary to establish a valid trend. This might be achieved by considering monthly instead of yearly accident data. Thus, 3 years of data would yield 36 observations. This number will be sufficiently large to test for a trend, if the seasonal variations are not too large. This procedure, however, would demand a fairly large number of accidents.

By adding several more before and after observations, the authors have constructed what is called an interrupted time series design. The use of this design in the North Atherton Street case would allow an evaluator to control for maturation by measuring the trend and the cycles. As an added benefit, potential difficulties arising from statistical regression can also be estimated.

The final threat to the internal validity of a before and after study is history. The results of the study are rendered uninterpretable if some extraneous event that might be plausibly related to a reduction in accidents were to occur between the two observations. In this case, the completion of any one of several other local transportation projects might reasonably affect the accident count or rate on North Atherton Street. Increased police surveil-

lance along the site, as requested by local residents, might reduce speeds and, thereby, the number of accidents.

Controlling for the potential threat of history necessitates the addition of a comparison site to the design. The comparison site must be carefully chosen so that it reflects the effects of historical events in a way similar to that of events at the project site. In the present example, sections of North Atherton Street on either side of the project site might well be appropriate. Accident data for a comparable period of time would also have to be collected for each comparison section.

The reader will note that the possibility for a selection bias has been introduced along with the comparison sites. In what ways do the sections on either side of the project site differ from the project site? One difference is that the average daily traffic counts are smaller, suggesting that accident rates rather than counts might be appropriate. Other differences will probably be apparent to the district traffic engineers. If it is deemed possible to relate these differences to accident rates, then appropriate evaluation design responses will be in order.

The before and after design is also subject to the treatment by setting and treatment by selection threats to external validity. It was suggested earlier that the most effective strategy for handling these threats was replication. Safety improvement projects are replicated frequently and the evaluation results should be analyzed according to project site type and accident type. Such analyses are not difficult given the available data base, and they would provide valuable information about those situations for which particular countermeasures are particularly effective.

Evaluations of typical safety improvement projects also suffer from problems with multiple treatment interactions. As this project has been described, there are at least four countermeasures being implemented in different combinations along the project site. It is unlikely that this particular mix of countermeasures will be applied to any other site. If information that can be generalized is to be gleaned from the North Atherton experience, then the countermeasure applications themselves must be able to be generalized.

One way to address this issue is to consider the North Atherton Street project as several projects instead of a single project. The site may be divided geographically into sections to which only one, two, or three countermeasures are being applied. Defining several project sites in this way results in what is called a multiple treatment design. The requisite information may be easily collected from available collision diagrams. The results from a multiple treatment design are more useful because single as well as the interactive treatment effects will be available. Additional analyses by accident type will make the results more valuable. These latter analyses serve as an explicit test of the expectations stated by the traffic engineers. By partitioning the available data set according to treatment type and accident type, the maximum amount of information can be gained from the North Atherton Street project.

The Ridesharing Program in the Delaware Valley Region

The ridesharing program was originally developed in response to the oil embargo of 1973. The Emergency Highway Conservation Act of 1974 authorized localities to use federal-aid highway funds to implement services encouraging the use of high-occupancy vehicles (carpools and vanpools). Later emphasis on transportation system management has stimulated renewed interest in the program and has made it an integral part of the transportation improvement program. Ridesharing program activities originally included information and matching procedures, service to major employers and individual applicants, resolution of institutional and legal barriers, and promotion by mass media.

The Delaware Valley Regional Planning Commission (DVRPC) is the metropolitan planning organization (MPO) in charge of ridesharing activities in the Philadelphia region. In the 3-year period between 1974 and 1977, it received $360,000 for project activities. At the end of the period, the FHWA requested all implementing agencies to submit a report documenting both program activities and degrees of participation. The DVRPC spent approximately $30,000 and compiled a 158-page evaluation document (8) summarizing the events, accomplishments, impacts, and current status of the program administered in the Pennsylvania counties within the MPO. A consulting firm was retained to assist DVRPC's staff in the development and administration of several surveys. Following the program's inception, energy supplies became less erratic. As a consequence, many ridesharing projects were abandoned. Subsequently, the scope of the program narrowed to implementation of employer-based ridesharing programs. Concomitantly, the evaluation requirements were substantially reduced.

In 1978 DVRPC received an increased 3-year contract under Urban Systems funds and was required by PennDOT and local county governments to show cost-effectiveness of the program to justify continued funding. During that period, DVRPC produced annual closing reports. The 1979-1980 report will be reviewed here. This type of evaluation study can be categorized as ex post facto. The 46-page evaluation report (9) describes implementation activities for the contract period.

The program was assessed as being successful on the basis of the 38 new companies processed and the 107 newly formed vanpools in the 15-month program period. These figures represent a substantial increase when compared with the period between 1974 and 1977. The mass media marketing campaign is claimed to have resulted in an increased number of inquiries during the contract period. The last and perhaps the most heavily relied on measure of effectiveness used in the report is the estimate of travel cost savings. These savings were calculated to be $1.3 million annually, yielding a benefit/cost (B/C) ratio of 8/1 (average annual project cost was $170,000). This savings figure does not include the indirect benefits of reducing pollution and congestion or the cumulative benefits of operations from previous years.

Implicit in these calculations is the assumption that the 107 new vanpools and about 850 new carpools were formed as a result of project activities. The estimation methodology suggested by the U.S. Department of Energy and the Pennsylvania Governor's Energy Council also assumes that the riders are diverted from driving single occupancy vehicles (occupancy rate 1.3), that the car left at home is used an average of 1.7 miles per day, and that the vanpools or carpools once formed remained in operation for the remainder of the contract period.

The evaluators admit failure to control the relevant threats to internal validity for history. External factors did occur during the contract period and may have contributed to an increased interest in ridesharing. Such factors include the 100 percent increase in gasoline prices during 1979. However, management's motivation to combat absenteeism or

tardiness resulting from gasoline supply interruption, to substitute noncash benefits in lieu of salary increases, and to improve employee morale may have contributed to participation quite independently of the program.

Besides the history and selection related problems, another threat that might be called treatment concretion is apparent. As the ridesharing program matured, institutional problems were ironed out and the initial inertia was overcome. At the same time companies and individuals became more willing and open to the idea of sharing a ride. There is a tendency in corporate management to follow the example of other successful programs, ridesharing included. Significantly, about one-half of the new vans were based at companies where original pools were established between 1974 and 1977, suggesting that existing vanpools acted as internal advertisement, which promoted new vanpools.

Another problem with attributing an increase in ridesharing to the program itself is that this is a service program, which by its very nature, has to respond to demand. Therefore the growth of the program may be the effect rather than the cause of spreading ridesharing. That service aspect of the program makes any advance planning or program innovation difficult, and, therefore, the usefulness of feedback information becomes limited. For example, a month-long strike by transit employees tied up the entire DVRPC staff in providing emergency matching assistance to commuters.

Although no attempt is made in the report to generalize from past experience to other geographic areas, there is a tendency to draw inferences on the basis of experiences with the first large companies to participate. The companies that willingly decided to promote ridesharing during energy shortages are not directly comparable to other companies of different sizes under different energy supply conditions. Unfortunately the format of the report might lead ridesharing project operators from other cities to believe that they can transfer the Philadelphia experiences to their own situations.

One major problem not addressed in the report is that measurement of the effectiveness or success of the project does not reflect the original intention of the program. The ultimate objective of the ridesharing project is to save energy by encouraging single occupancy automobile drivers to change their style of commuting. A survey to determine whether or not such changes have resulted in net energy savings and what happens to the car left at home would be more informative than the number of new carpools or vanpools formed.

One criticism that has been made of the program by member governments of the DVRPC Board is that ridesharing promotion is attracting commuters from the suburban rail system. The recent sharp increase (doubling in the past 2 years) in transit fares has made vanpooling and carpooling an attractive alternative for commuters. As a result, some of the suburban counties are faced with the threat of a reduced commuter rail service because of a loss in ridership. Another reason for opposition to the program is that the counties want to get a share of DVRPC's funds and they want to participate in the program. In recent years there has been a shift toward greater involvement of counties in paratransit brokerage services.

A final point to be raised concerns the quality of performance in estimating the impacts of ridesharing. The evaluators readily admit that there is a high degree of uncertainty in their estimates; however, they justify them by stating that the additional surveys necessary to obtain more reliable estimates would be more costly than the project budget allows. The project manager determines the amount of effort, content, and methods to be used in monitoring the program, subject to the approval of the DVRPC Board. Evaluation does not usually exceed 5 percent of the project's budget; and from the point of view of the implementing agency, the closing report fulfills a requirement and provides an opportunity to demonstrate the need for continuation of funding. There is no evident desire on the part of board members to increase the research and evaluation allocations.

A substantial portion of the evaluation report is devoted to recommending the maintenance and expansion of existing services and programs to serve the growing demand. The 1981 closing report, entitled "A Policy Paper," devotes most of its discussion to proposals for program continuation (10). Incidentally, during 1982 the ridesharing budget will be almost doubled from existing levels with the addition of planning and federal aid funds. The rationale is that ridesharing is both an alternative to those portions of commuter rail service that are inefficient and a method for achieving the 1982 air quality standards. Although budget cuts are affecting most other transportation programs and staff are being laid off at DVRPC, ridesharing is increasing its budget and its staff. This year, for the first time, PennDOT has developed a set of guidelines for program process review.

The U.S. Department of Transportation (DOT) realizes the importance of most of the problems of evaluation that are addressed here. Wagner (11) in a study prepared for the FHWA, reported significant deficiencies in the 1977 (original) evaluation methodologies used by agencies across the nation. For example, sample results were extrapolated to the wrong populations, and frequently extra travel mileage attributable to carpooling was not considered. The information in the report suggests an experimental plan that stratifies the population into segments that are either exposed or unexposed to different elements of carpool matching and promotion activities. The purpose of the plan was to determine (a) the net changes in commuter travel modes caused by project activities and (b) the portion of the changes in commuter travel attributable to different types of project activities.

In 1979 the FHWA staged a 2-year National Ridesharing Demonstration Project. The objectives of this program were to implement and evaluate (cost-effectiveness of measures and effectiveness of organization) comprehensive and innovative approaches to ridesharing at 17 sites that are representative of various geographic areas and population sizes. The Transportation Systems Center of DOT has been assigned the responsibility for the evaluation. The data collection will be performed by the grantees. DVRPC is one of the agencies chosen.

The evaluation will consist of a detailed documentation of each case, a cross-cutting comparative analysis of all projects on common technical and institutional problems and special studies at a few sites to address travel behavior research questions (12). The evaluation design proposed in the FHWA effort appears to address the most critical questions.

If the objectives of the ridesharing evaluation are to provide an account of what happened, convince grantors that there is a need for a continuation of the program, and show that it meets the criteria (B/C ratio) set by PennDOT and the local governments, the effort invested by DVRPC in evaluating the program has been successful. The fact that the budget is increased every year justifies this conclusion. They have learned from experience that they can be most effective by selling the business com-

munity on the program and they have concentrated their efforts to that end. Common sense, good judgment, effective public relations, and sensitivity to political realities are requisites for selling the idea to the public and for convincing funding agencies of the need for additional funding in the future. However, if the objective of the evaluation is to explain the effects of the program on travel behavior and to be able to use this knowledge for future decisions, it has failed. In this era of conflicting interests and limited budgets, expansion of ridesharing means that some other program will be cut or some area or group of people will be adversely affected. It appears, therefore, that there is an increased need for a more conclusive effectiveness evaluation. Some hard decisions will have to be made about where the money for improved evaluations and monitoring will come from. It may be advisable to assign the task of evaluation to an independent party to avoid any possible bias.

The Promotional Reduced Fare Program in Scranton, Pennsylvania

The third project is the promotional reduced fare (PRF) program being conducted in Scranton, Pennsylvania, by the County of Lackawanna Transit System (COLTS). Primary funding for the project is provided by an UMTA Service and Methods Demonstration (SMD) program grant. The total cost of the program is estimated at $235,671; approximately one-half of this amount is being used to cover lost revenues and roughly one-third is being used to cover planning and evaluation costs. Eighty percent of the funding is being provided through the SMD grant; the remainder is provided by COLTS. The Transportation Systems Center is responsible for the technical direction and supervision evaluation, and the responsibility for conducting the evaluation is being subcontracted to Vista Systems of Princeton, New Jersey (13).

The UMTA SMD program addresses the need to improve public transportation services by sponsoring demonstrations of innovative transportation services and management techniques. The SMD projects focus on innovations with applicability to a diversity of settings and operations. The PRF program is under the SMD Pricing and Service Improvement Program (14).

The objectives of the PRF program are to determine the range of the effects of experimental transit marketing policies. Instead of being interested primarily in the effectiveness of the program in attaining local goals, UMTA's interest in the program is in the ability to generalize the findings to transit agencies around the nation.

The PRF program was implemented in January 1981 to determine the effects of 1-month fare reductions on transit ridership and public attitudes toward transit. In addition to the effects on total ridership, the program was designed to determine

1. The market segments of prior riders who increased their rate of transit use because of the fare reductions;
2. The market segments of new riders who began using the transit system because of the fare reductions;
3. The portion of the new ridership generated by the fare reductions that is retained after the reinstatement of the regular fare; and
4. The actual costs and future revenue benefits of the promotions.

The PRF program consists of reducing the 45-cent regular base fare to 20 cents, 5 cents, and free fare for 1-month periods, with 5 months of regular fare between each promotion. The evaluation design consists of two components: an interrupted time-series design and an equivalent-time sample design with multiple, related (not identical) treatments. Historical ridership and revenue data, which is already being collected by COLTS on a daily basis, will be analyzed to show the effects of the reduced fares on aggregate ridership levels. Monthly checks on schedule adherence and crowding will be made to assess the operational impacts of the reduced fares.

The collection of data on historical trends rules out several of the threats to the internal validity of the evaluation. Inspection of the historical trend data can ensure that the fare reductions are not implemented at a time when ridership levels are unreasonably low, controlling for any regression threat. The long-term ridership data (from the previous year, for example) can be analyzed to account for any seasonal or increasing trends, accounting for the possible maturation threats. Finally, there is no reason to suspect the presence of an instrumentation effect.

The main threat to the internal validity of this design is the possibility that factors other than the reduced fare program may have caused any observed increases in ridership. Because no control group is present, the occurrence of any event that could influence ridership levels concurrent with the implementation of reduced fares would seriously weaken the interpretability of the results. It should be noted that consideration is being given to using the city of Wilkes-Barre, Pennsylvania, as a control site.

Two factors may help moderate, though not eliminate, this possible history threat. First, a specific temporal pattern of the ridership response to the reduced fares was hypothesized before the start of the demonstration program. It was hypothesized that a large increase in ridership would occur immediately when the reduced fare was introduced and that this ridership level would decline gradually as the novelty of the new fare tapered off. It was expected that a sharp decrease in ridership would occur when the regular fare was reinstated and that ridership would then stabilize at a level higher than the ridership level before the reduced fare month (reflecting the retention of new riders and new trips). If this response pattern is observed, the inference that the fare program is responsible for the ridership increase is strengthened. This specification of the expected response is encouraged when a time-series design is employed, particularly when the response to an intervention is expected to be delayed rather than immediate.

Second, if a rival event is present, the three fare reduction months can be regarded as repetitions of a class of treatments (fare reductions) and inferences can be drawn as to whether fare reductions increase ridership. However, no inferences could be made regarding the differential effects of the varying levels of fare reduction.

Five on-board surveys and three telephone surveys are being conducted to collect disaggregate information on the number of new riders attracted by the reduced fare months, new trips generated, and market segments attracted by the promotions. Because this aspect of the demonstration evaluation is oriented toward answering more specific, exploratory issues rather than the effectiveness of the program, this part of the design will not be described in detail here. However, two points should be mentioned. First, this additional data collected from individuals will provide more information on the particular influences of the promotions and will help determine whether extraneous influences are present. Second,

the collection and analysis of this survey and interview data account for a large proportion of the evaluation costs.

Although the surveys may provide valuable information on the nature of any changes observed, the effectiveness of the program can be determined without incurring the cost of surveys and interviews. Although answers to these exploratory questions are necessary in federally funded demonstration projects, adequate archival data for determining program effectiveness, such as ridership and revenue figures, are frequently available at little marginal cost.

Two factors lend strength to the external validity of the design. First, the design could be regarded as a replication or refinement of numerous free fare and fare reduction programs conducted both formally by UMTA (15) and as marketing programs conducted by transit properties across the nation. The rigor of the evaluations of these programs varies dramatically; however, viewing them as a whole can provide indications of the extent to which the findings of the PFR program can be generalized. Granted, this question cannot be answered conclusively, but this background of experience can provide important information.

Second, the random sampling procedures used for selecting the respondents for the telephone surveys increase the likelihood that the samples are representative of the given population. Although the population sampled (and treated, for that matter) is still specific to one setting (only one city, Scranton, Pennsylvania, is included), for practical purposes this random sampling procedure is the optimal route for ensuring that the results will have general application. Nevertheless, the critical reader of the study must still ask, How representative is Scranton of all cities (or medium-sized cities, etc.) and how representative is COLTS of all transit systems?

One remaining problem involves an aspect of the program not mentioned earlier. After the first fare reduction was implemented and the regular fares were reinstated, the decision was made that the retention of new riders was not high enough to ensure that the revenues lost during the fare reduction period would be recovered within the same fiscal year; this was an objective regarded as highly desirable by all the parties involved. Therefore, implementation of an employer-subsidized pass program on a continuing basis is being considered as an attempt to increase the number of new riders retained after the reinstatement of the regular fare. This pass-subsidization program can be regarded as an additional treatment affecting some of the new and prior riders.

Therefore, some of the information on the effectiveness of fare reductions alone in increasing and retaining ridership may be lost (i.e., it may be difficult to determine what the effects on ridership would have been if the pass-subsidization program had not been implemented). The introduction of the pass-subsidization program will not, however, make the results of the evaluation completely uninterpretable. The pass-subsidization program will affect only one segment of the ridership: the riders working for employers who are participating in the pass-subsidization program. Because the PRF program is collecting data from individuals (instead of aggregate ridership data), information on the effects of the program on market segments not affected by the pass-subsidization program will still be available. Also, the telephone surveys can be tailored so that they provide some information on the additional effects of the pass-subsidization program on travel mode choice.

The PRF program evaluation illustrates the problems and the benefits of conducting studies in a dynamic environment. Because of the many known and unknown factors that may be influencing any given behavior or occurrence, attributing any change to a single cause is difficult in an open setting. Even when an evaluation is designed so the influence of many possible extraneous factors is controlled, there are frequently numerous constraints, including financial and political constraints, that may be more important than the particular questions being addressed and that must be considered by the decision makers.

The implementation of the pass-subsidization program is such a case. COLTS and UMTA realized that an important objective, the recovery of lost revenues within the same fiscal year, would not be attained as the program existed; therefore, they fine-tuned the program to try to meet this objective. The management can hardly be faulted for making changes to improve the cost-effectiveness of the program's operation. The potential benefits of the pass-subsidization program far outweigh the cost of the loss of some of the information, particularly in today's financial and political climate. It should not be overlooked, however, that it was a well-designed evaluation that provided the management of the PRF program the information necessary to suggest the need for this fine-tuning.

SUMMARY AND EXTENSION

The primary focus thus far has been the experimental plan. Valid experimental plans are crucial because they provide an evaluator with the basis for inferring a causal relationship between a project or program and an observed difference between MOEs. Critical assumptions permitted by uses of random selection and assignment are more difficult to justify in nonexperimental situations. The necessary conditions may be approximated by explicitly controlling threats to internal and external validity. Six classes of threats to the internal validity were described. Experimental plans that control these threats generally take the form of either nonequivalent control groups or interrupted time-series designs. Replication was suggested as the best way to deal with threats to external validity.

In the review of the three case studies that followed, it became obvious that there is an unnecessarily wide gap between capability and practice in the design and conduct of evaluation studies. The authors do not claim that findings from this small sample of case studies can be generalized to all evaluations. The intention was to demonstrate the implications of using a strong evaluation plan in selected cases and to place them in perspective with the implementing agency's objectives.

The first two evaluation plans, consisting of a before and after design and an ex post facto design, were found to be deficient thus rendering the results of little use either to the implementing agency or to the decision makers. In the case of the safety improvement project, maturation and history were found to be the most serious and probable threats to the design's internal validity, and multiple treatment interaction was judged to be a serious threat to its external validity. In response a multiple treatment, interrupted time-series design with a nonequivalent control group was recommended.

Although the reports of the ridesharing program are successful in meeting the agency objectives of continuation of the program, they fail to address the threats of history, maturation, selection, and interaction. Furthermore, they do not attempt to generalize to other situations. The Wagner report

and the new FHWA initiatives do propose experimental designs that may be able to address most of these issues.

The experimental plan associated with the promotional fare reduction program is considerably stronger, and several important points are implicit in the example. The explicit a priori statement of hypotheses, or expected results, strengthened the evaluation. This strength is achieved if the hypotheses are supported because of the existence of a much smaller set of rival hypotheses that predict identical results. In addition, intermediate results from the evaluation provided valuable feedback to decision makers that allowed for desirable programmatic adjustments. The implementation of these adjustments, however, illustrates the sensitivity of experimental plans to changes in the intervention. These projects and their associated evaluation plans, threats to validity, and suggested remedies are summarized in Table 1.

In general, the ideal of a completely valid effectiveness evaluation is unattainable. It is clear, however, that relatively little effort is required to strengthen most experimental plans substantially. Given that the results of a single study will never be entirely conclusive, the evaluation has certain obligations. The interpretation and utility of a given study require that the evaluator state explicitly both the assumptions underlying the experimental plan and the validity threats that cannot be effectively eliminated.

The experimental plan has been discussed here independently of the other components in an effectiveness evaluation. This picture is not completely accurate because all the components are interdependent. The selection of an appropriate MOE, the methods by which data are collected, and the appropriate application of statistical tests are also important components of an effectiveness evaluation. Each of these areas also has a gap between theory and practice. For example, a heavier reliance on attitudinal or nonarchival, unobtrusive MOEs could alleviate frequent difficulties experienced between sampling models and sample sizes. Another potential area for improvement is the integration and more efficient use of existing data bases. The reader is directed to Highway Safety Research Center (2) and Goodell-Grivas, Inc. (3) for discussions of these related issues.

The development of a strong evaluative capability depends on more than the existence of the requisite technical capability. Some of the important institutional, financial, and attitudinal barriers that inhibit institutionalizing this capability are addressed in the following paragraphs.

The conduct of valid effectiveness evaluations is expensive, both in time and money. For this reason, the issue of when to evaluate deserves consideration. Effectiveness evaluations are quite often desirable for two types of projects. These include controversial projects and demonstration projects or, more generally, any project in which the impacts are uncertain.

As Quade (16) points out, "A demonstration project is by its nature an experiment; and unless the results can be evaluated properly, all that may be demonstrated is that it is possible to spend public funds in a particular way." Many demonstration projects are currently evaluated, and generally quite well, on the national level through UMTA's Office of Service and Methods Demonstration. Given the funding mood of the current administration, it is probable that some portion of the financial responsibility for evaluating innovation and demonstration projects will be passed to the states. The possibility of adopting the federal SMD model on the state level merits further examination.

The second category of projects for which evaluations are often desirable is the controversial ones. These include projects with conflicting objectives or projects with potentially adverse side effects. For example, although the addition of an extra lane for left turns on North Atherton Street in State College might reduce vehicle accidents, some residents believe that it might create an unsafe condition for the pedestrians crossing the street. Similarly the expansion of the ridesharing program in the Delaware Valley attracts commuters from the local rail service that is in opposition to the interests of saving the commuter lines. In such adversary situations it is often advisable to assign a third party the task of evaluating the projects. In the past there have often been sufficient funds to satisfy conflicting needs and interests; however, this is no longer the case. The equitable resolution of such conflicts necessitates an increased accountability of the expenditure of public monies. Valid effectiveness evaluations are the preferred method for assessing this accountability.

Accountability is closely linked to objectivity and believability within the implementing or evaluating agency. Assumptions and results may unduly favor one perspective over another. This lack of objectivity is often the result when agencies are forced to evaluate their own work for the purpose of securing additional funding. On the other hand, the evaluation may be strengthened by integrating the planning and evaluating functions. The administrator's dilemma of how to allocate funds can be solved by segregating the funding sources of planning and implementation from those of evaluation.

In addition, the prevalent administrative atmosphere does not appear conducive to the productive integration of social experimentation with more traditional decision making modes. Because funding decisions are usually separated from implementation decisions, administrators are locked into support for their particular project for fear of losing funds or their jobs. A solution has been offered by Campbell (17). He suggests that administrators shift

TABLE 1 Summary of Case Studies

Type of Project	Implementing or Evaluating Agency	Evaluating Agency Objectives	Experimental Plan	Major Threats to Validity	Suggested Remedies
Highway safety improvement	State district office	Fulfill federal requirement	Before and after	History, maturation, regression, instrumentation, treatment interaction	Multiple treatment, interrupted time series with a nonequivalent control group
Ridesharing	Metropolitan planning organization	Fulfill state and local requirements and secure continuation of funding	Ex post facto	History, maturation, selection, interaction with selection by history, treatment concretion, external validity	Multiple treatment with a nonequivalent control group, comparative case study analysis (redefine MOEs)
Promotional reduced fare	Transit authority and federal research center	Evaluate innovative ideas and conduct exploratory research	Interrupted time series with multiple related treatments	External validity	Reapply treatment

from advocating a specific program to advocating the seriousness of the problem being addressed and offering their recommended solution as the best of several alternatives.

Given the trend at the federal level toward a reduction of funds for transportation projects and toward a transfer of power and responsibility to the states, there is a need to improve planning and research. The wise use of past experience through the regular application of effectiveness evaluations will contribute to this improvement. A firmer understanding of the nature of experimental plans by both transportation professionals and other administrative decision makers is a necessary step. Evaluation funds, wisely spent, may be quickly recovered in the form of improved decisions. In the coming years states will have to assume a greater initiative in the development of new corrective and creative programs. Effectiveness evaluations of key programs can provide states with the information necessary to tailor these programs to their own specific needs.

ACKNOWLEDGMENT

The authors wish to thank Mark D. Gurtler, a graduate assistant in the Division of Man-Environment Relations at The Pennsylvania State University for his work on the project. The authors also thank the many transportation professionals, particularly those at PennDOT, who were gracious with their time and tolerant of repeated visits and telephone calls. This study was supported by a grant from the Regional Analysis Center of the Institute for Research on Land and Water Resources at The Pennsylvania State University.

REFERENCES

1. The Evaluation of Highway Safety Programs: A Manual for Managers. DOT HS 802 525. National Highway Traffic Safety Administration, U.S. Department of Transportation, Feb. 1979.
2. Accident Research Manual. Highway Safety Research Center, FHWA, U.S. Department of Transportation; University of North Carolina, Chapel Hill, Jan. 1980.
3. Goodell-Grivas, Inc. Highway Safety Evaluation Procedural Guide. FHWA, U.S. Department of Transportation; National Highway Institute, Southfield, Mich., March 1981.
4. T.D. Cook and D.T. Campbell. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago, 1979.
5. The 1980 Highway Safety Standard Report: Report of the Secretary of Transportation to the United States Congress. Office of Highway Safety, FHWA, U.S. Department of Transportation, July 1980.
6. 1980 Annual Report. Pennsylvania Highway Safety Improvement Program, Traffic Engineering and Operations Division, Bureau of Highway Services, Pennsylvania Department of Transportation, Harrisburg, Sept. 1980.
7. P. Loukissas and J. Mace. Conversion of Pennsylvania Accident Data to Account for Change in Reporting. Pennsylvania Department of Transportation; The Pennsylvania Transportation Institute, The Pennsylvania State University, University Park, Pa., 1982.
8. Emergency Highway Energy Conservation Program Ridesharing in the DVRPC Region 6/74-6/77 Closing Report. Delaware Valley Regional Planning Commission, Philadelphia, Pa., 1978.
9. Ridesharing Implementation Program Pennsylvania Portion of the Delaware Valley Region, April 1979-June 1980, Evaluation Report. Delaware Valley Regional Planning Commission, Philadelphia, Pa., July 1980.
10. The Third Mode-Ridesharing in the DVRPC Region 1981 and Beyond--A Policy Paper. Delaware Regional Planning Commission, Philadelphia, Pa., May 1981.
11. F.A. Wagner. Evaluation of Carpool Demonstration Projects, Phase I Report. DOT-FH-11-9269. FHWA, U.S. Department of Transportation, Aug. 1978.
12. National Ridesharing Demonstration Program Evaluation Plan. Transportation Systems Center, U.S. Department of Transportation, Princeton, N.J., Jan. 1980.
13. Vista Systems Inc. Evaluation Plan for the Promotional Reduced Fare Demonstration in Scranton, Pa. Transportation Systems Center, U.S. Department of Transportation, Princeton, N.J., July 1981.
14. Service and Methods Demonstrations Program Annual Report. UMTA, U.S. Department of Transportation; National Technical Information Service, Springfield, Va., July 1978.
15. De Leuw, Cather and Co. Findings of Preliminary Analyses of the Trenton, New Jersey Off-Peak Fare-Free Transit Demonstration. UMTA/TSC Project Evaluation Series, Report UMTA-NJ-52-0001-78-1. Transportation Systems Center, U.S. Department of Transportation, Princeton, N.J., Jan. 1979.
16. E.S. Quade. Analysis for Public Decisions. Elsevier, New York, 1975.
17. D.T. Campbell. Reforms as Experiments. American Psychologist, Vol. 24, No. 4, April 1969.

Publication of this paper sponsored by Committee on Social, Economic and Environmental Factors of Transportation.