Evaluation Research: Some Problems and a Possible Solution

William E. Alexander and Joseph P. Farrell
Ontario Institute for Studies in Education

Much evaluation research purports to identify the goals of a program and assess the degree to which these goals are being achieved. It is the position of this paper that usually goals are not identified by the researcher but, rather, are created by the researcher. It is argued that this process increases the probability that negative evaluation findings will be ignored, and creates potentially serious difficulties when findings are taken seriously. Moreover, the typical process of educational evaluation research has the potential (a) to usurp the authority of legitimate policy makers by permitting researchers to choose program goals, and (b) to impose a homogeneous set of values on a number of programs even when these programs have a legitimate claim to diverse sets of values. In response to these and other problems, the authors developed an alternative evaluation design. This technique is described, and some of the results of its first application in an evaluation of the effectiveness of student governments in 37 secondary schools in Ontario are presented.

Dans le domaine de l'éducation, la plupart des recherches en évaluation ont pour but de mettre en évidence les objectifs d'un programme et d'en évaluer le degré de réalisation. Les auteurs de cet article prétendent que, en règle générale, ces objectifs ne sont pas identifiés mais bien créés par le chercheur. Ces mêmes auteurs soutiennent que cette démarche accroît les risques de voir les conclusions négatives négligées ou de créer des difficultés potentiellement graves lorsque les conclusions de la recherche sont prises au sérieux. Cette démarche habituelle de la recherche en évaluation risque en outre (a) d'usurper le pouvoir légitime des concepteurs de programmes en permettant aux chercheurs d'en déterminer eux-mêmes les objectifs et (b) d'imposer un système de valeurs homogène à un certain nombre de programmes alors que ceux-ci pourraient légitimement s'appuyer sur des systèmes de valeurs différents. Pour résoudre ces problèmes et plusieurs autres, les autres ont mis au point une nouvelle technique d'évaluation. On fait d'abord un exposé de cette technique; ensuite on présente les conclusions tirées d'une première application de la technique. Cette première application s'est faite à l'occasion d'une évaluation de l'efficacité des conseils étudiants dans 37 écoles secondaires de l'Ontario.

Throughout the world, the decade of the 1960s was an era of spectacular quantitative growth in educational systems. Hundreds of millions of dollars were invested in building more schools, training more teachers to staff them, and inventing and implementing new educational programs, either to better serve the clients of traditional schools or to serve those young people perceived to be unsuited for or poorly served by existing schools. As with most booms, the one in education slowed down as abruptly as it began. Questions regarding the internal efficiency of schooling came increasingly to be seen as important. How well were schools

serving their objectives, and how could they serve them better at the same or lower cost? This change in emphasis has been seen throughout the world, but has been particularly evident in richer nations. In the United States accountability has become a movement, a national assessment has been mounted, and evaluation of educational programs has become a major sub-industry. In Canada, too, although more quietly, and less enthusiastically institutionalized, the trend toward slowing quantitative expansion and evaluating internal efficiency is evident, as taxpayers and their representatives are asking hard questions about the value of educational policies and programs. Thus, although evaluation studies in education have long been with us, they have taken on a new prominence and salience during the past few years.

It is in this context that we wish in this paper to (a) discuss some serious problems we see in the way most evaluation research in education (and indeed in the social policy field generally) is conducted, and to (b) present a research strategy we believe may avoid the difficulties which concern us.

Precisely because evaluation of educational programs is such a broad and varied field of activity, neither the criticisms discussed below, nor any other set of criticisms, could be expected to apply with equal force to all evaluation studies. We argue that all of our criticisms apply to most educational evaluation work, and that there are few evaluation studies to which at least some will not apply. Our arguments, as a set, will be at least applicable to those cases where an evaluation is built in at the beginning of an experimental or pilot program and to those rare cases where specific program objectives and ways of measuring them are specified by legislation. They will be most salient to those more typical cases where an evaluation is requested of an already existing program or activity whose goals are not (and never have been) well delineated, and which is widespread, meaning that any change will require the active collaboration of a large number of individuals.

In the area with which we are dealing there is much confusion regarding what labels to apply to various kinds of situations. It is increasingly common to distinguish between evaluation and research. In this view, evaluation refers to activities designed specifically to determine how well a program or activity is functioning, or if it is achieving its objectives, whose results are normally directed to individuals responsible for the program or activity. Research, in this usage, refers to activities designed to increase our knowledge of some phenomenon (with the emphasis on how rather than how well), usually following some established paradigm (explicit hypothesis testing, the use of accepted data collection designs and statistical methodology, etc.,) whose audience is a general scholarly community. In thinking about the problems discussed below we have not found this distinction helpful. Evaluation activities normally follow accepted research paradigms (the implicit hypothesis is typically that acti-

vity X produces effect Y, and standard procedures for collecting and analyzing data are usually employed). Conversely, much important evaluation information is the result of straight research. Policy decisions and aggregated changes in the behavior of individual teachers and administrators are more commonly the product of cumulated research results than of particular evaluation exercises. We prefer to think of evaluation activities as a subset of research, and thus throughout the following pages will use the term "evaluation research."

However, when researchers undertake specific evaluation exercises, there are two characteristics of the situation which they must bear in mind: (a) There is a clearly specified client — someone asks for the evaluation — who normally has legitimate responsibility for the program or activity being evaluated, and (b) It is expected that the results, if they are in any sense negative, will stimulate behavior change in the (usually many) individuals who directly operate the program or activity — administrators, teachers, students, etc. It is in considering the constraints these characteristics place on the behavior of the researcher-evaluator that the problems discussed below arise.

PROBLEMS

The aspects of evaluation research with which we shall deal are, first, the ways in which evaluation researchers typically choose the goals or objectives in terms of which a program or policy is to be evaluated and, second, the ways in which the results of evaluation research are dealt with by educational systems. These two concerns are interrelated.

How Goals are Chosen

It should be obvious that, as Suchman (1967) has noted, evaluation research begins with "the identification of goals to be evaluated" (p. 51). One can hardly determine whether a program or policy is meeting its objectives if one cannot specify what those objectives are. However, although the necessity of identifying program or policy goals is obvious, the act of so identifying them is extraordinarily difficult. Indeed, it is so difficult that it is seldom done. We would argue that goals, as used in evaluation research, are not typically identified by the researcher; rather they are created by the researcher.

If one seeks to identify the goals of an educational program or policy, one would ordinarily turn either to official policy documents or to the statements of individuals who have either created or are implementing the policy or program. Yet the evaluation researcher who turns to either of these sources in search of goal statements which can be operationalized will seldom be satisfied. In official documents one seldom finds explicit statements of goals (or aims or objectives). Moreover, if one does encounter statements of goals or aims in official documents they are ordi-

narily so vague as to be useless for the evaluation researcher. How does one evaluate whether a program or policy (or an entire system) is achieving the goals of "creating good citizens," or "meeting the needs of every child," or "fostering a zest for learning," or "encouraging adaptability to a changing world," or "providing a sound basic education"?[1]

Many theorists suggest that this reigning vagueness in official documents is of little importance, for official goals differ in any case from real goals — the goals held by the actors in a system — and it is upon the latter that one should focus evaluation studies (see, for example, Buck, 1966, p. 109; Perrow, 1961, p. 853; Etzioni, 1964, p. 7; for a recent review of the literature regarding this issue, see Hall, 1975, pp. 13-17). This advice is of little help to the evaluation researcher, however, for it is equally difficult to abstract goals which can be operationalized from the statements of individuals within an educational system. If one can get an answer at all to a question about goals put to an educational official, the statements tend to be as vague as those encountered in official documents. Moreover, the statements made by a single individual will often vary over time. Beyond this, even if one can get useful goal statements from one individual in a system or program, it is very rare to find another actor in the system who will provide a similar set of goals statements — and even if there is some overlap among various actors' lists, the common goals will not necessarily have the same priority ranking for each individual. In complex educational systems, where decision-making structures are pluralistic, fluid, and often so poorly defined as to be opaque (cf. Alexander & Farrell, 1975, chap. 2), it is difficult to identify a single individual whose personal goals can be taken to represent the organization's goals.[2]

What the arguments above suggest, in summary, is that statements about specific goals are simply not part of the lexicon of most educational organizations as they go about their daily business, or even when they undertake their most serious business, such as policy formation. This is not a problem peculiar to educational systems; it is endemic to social organizations, which, as Gross (1965) has noted, can be characterized as providing "intangible end-product services . . . on a non-sale basis to an intangible, unorganized or reluctant clientele" (p. 205). Indeed, long experience in wrestling with the problem of identifying organizational goals has led many social scientists to a position bordering on despair. In a recent article, Lawrence Mohr (1973) begins his argument by noting that "Much has been written in recent years about the concept of organizational goals, some of it rather discouraging; the most frequently cited papers offer little hope that the concept will have any real utility for social scientists" (p. 470).

What, then, are evaluation researchers to do? They must have goals in terms of which to evaluate performance, yet they are most unlikely to get useful goal statements from either policy documents or actors in the system. We suggest that what ordinarily occurs is that evaluation researchers

use official policy statements, conversations with actors in the system, and their own intuitions about educational systems as general guidelines for deciding which broad classes of outcomes will be measured. These guidelines restrict but do not determine the selection of specific measurable goals or outcomes. It is useful to note that this is more a process of induction than of deduction. As George Homans (1961) has noted, while deduction has definite rules, the rules of logic, induction "has no rules of procedure that will ensure you success" (p. 10). It is a creative act. The choice of specific outcomes to be measured is a creation of the researcher, and necessarily reflects his or her personal values. Gail Stewart (1972) has made the point more generally with reference to all social indicators: "Since every social indicator, no matter how apparently isolated, represents a way of viewing reality and hence represents a value judgement on the part of its creator, shouldn't every social indicator be signed by its creator as an artist might sign his work?" (p. 30).

We are suggesting, then, that the specific outcome measures, in terms of which performance is actually judged in evaluation research, are the private creations of evaluation researchers (although they may, of course, have borrowed them from the work of some other researcher), and as such necessarily have little to do with the individuals in the system being evaluated. Regarding such external creations the individuals in the system have two options: They may ignore the results, or they may take them seriously. Each option carries its own problems.

How Findings Are Dealt With

Ignoring findings

It is well known that negative findings tend to be resisted. Nobody likes bad news.[3] This is especially a problem for evaluation research in education, since evaluation of a particular program or policy easily shifts into evaluation of the individuals responsible for creating or administering it (even if an evaluation is not planned that way, it is certainly ordinarily perceived that way by actors in the system being evaluated). Much has been written regarding this problem, but one of the most salient points for our purposes is made by Reginald Carter (1971), who presents this proposition: "The greater the difference between the client's or manager's concept of the social reality being studied and the research findings, the greater will be his resistance to negative findings" (p. 118).

The process of creating goals for evaluation research which we have discussed above appears to maximize the probability that the "concept of social reality" embedded in an evaluation research will differ from that of key actors in the system being evaluated. Concretely, it makes it all too easy for an administrator (say, the principal of a school) to claim that the outcomes being evaluated are irrelevant to the real goals of his program,

or that while the general goal being considered is important, inappropriate measurements were applied, or that success criteria used by the evaluation researcher are too high. If evaluation research is to respond to Donald Campbell's (1974) recent call that "the methodology of evaluation research should include the reasons for this resistance and ways of overcoming it" (p. 28), then some less "creative" way of specifying goals in evaluation work needs to be developed.

Taking findings seriously

Although rejection of negative findings is a problem, in that the money spent on evaluation is wasted and a possible opportunity for individuals responsible for the evaluated program to learn and improve their performance is lost, one can at least claim that no serious social harm ordinarily results (except, of course, in cases where a program whose outcomes are harmful to clients is continued or even expanded). On the other hand, unfortunate consequences may, and often do, occur when evaluation outcome measures are taken seriously. If individuals know their performance is going to be evaluated in terms of one indicator, or a small set of indicators, there is an entirely reasonable tendency to operate the system so as to ensure a good score on the chosen indicator(s). At best, this tendency can skew the efforts of a multi-purpose system (and whatever else may be said about the goals of education, we can certainly agree that there are many of them) toward meeting a small set of objectives, ignoring all those which are not being measured. At worst, it can lead to literal falsification of the evidence.

Examples of the dysfunctionality of performance measures abound in the literature (cf. Campbell, 1974, pp. 20-30). Perhaps the most familiar example in education is the effect of externally administered achievement tests. If used to evaluate schools (or indeed teachers), even if only informally, tremendous pressures are created to teach only the subject matter included in the test, to "cool out" students who are likely to score low before the test is administered, so their scores will not bring the average down, and in some cases, where it is possible, to corrupt the testing process by teaching precise answers to questions known to be on the test. These tendencies have led Campbell (1974) to formulate a "law": "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p. 29).

We would also note that if the results of evaluation research — in which the goals, as operationalized, are the creation of the researcher — are taken seriously, the process tends to controvert some important social or cultural values in our society. Specifically, such a research process tends (a) to usurp the authority of legitimate policy-makers by permitting

researchers to choose their goal criteria; and (b) to impose a single set of values — in the form of instrumental or ultimate (operational) goals (and often goal priorities) — on units which have legitimate claim to diverse sets of values.

The usurpation of authority. Evaluation research is ordinarily carried out with a view to influencing policy — which is to say, influencing policy-makers. Indeed, much of it is solicited and paid for by policy-makers. By virtue of their position, policy-makers have a legitimate claim to act on behalf of a body politic — a claim which an evaluation researcher does not ordinarily have. Yet, as we have noted above, it is the researcher — who has no publicly legitimated authority to make such a decision — who typically determines the goals in terms of which a program or policy will be evaluated and, if the results are taken seriously, the direction in which the program or policy will move. In short, the researcher takes upon himself broad powers to determine which part of a complex reality is (or should be) most important to the representatives of the public.

To cite one example, a research project in which one of the authors participated was an effort, supported by a ministry of education, to assess the effectiveness of an innovation in secondary school programming in terms of the post-secondary experiences and attitudes of secondary school graduates. The research team originally generated 16 measures which it was felt reflected the general aims of the program, as stated in policy documents. However, the team also felt that collecting data on all 16 measures would produce a questionnaire so long that response rates would be very low. It was therefore decided to eliminate several of the measures, including, for example, indicators of need-achievement, self-concept, and feelings about others. The decision, as is typical in such studies, was made by the research team without consulting either the ministry personnel who contracted the study or the principals whose graduates were being evaluated. The research team, not policy-makers, decided that an outcome such as changes in one's feelings of efficacy was more important as a program goal than changes in one's feelings about other people.

Another study, in which one of the authors participated, was designed to evaluate the effectiveness of a series of innovations at the primary level. Here again, policy documents cited a number of very general reform objectives. Again, only a small sub-set of these outcomes was actually assessed in the evaluation study. In this case, ministry personnel were involved in making the decisions, but it was quite clear that the technical arguments of the researchers (e.g., "it is difficult to measure attitudes among primary-aged children," or "we can't afford to measure actual teaching behavior,") carried the day. This case illustrates a very important point. Given the technocratic ethos of our society, even when policy-makers are formally involved in the selection of goals to be measured, the

expert aura of researchers will often permit their views (frequently influenced by the niceties of research design, what kinds of data are most easily handled, what kinds of concerns will produce publishable articles) to prevail.

The imposition of homogeneity. It has long been recognized that individuals, organizations, and units of government can, within an overall framework of generally agreed principles and aims, pursue different goals, or at least place different priorities upon commonly agreed goals. Different universities and colleges, for example, frequently have quite different priorities for their programs. While often constrained by service or expenditure minima and/or maxima imposed by senior levels of government, individual communities have considerable freedom to determine which public services will have the highest priority within their domain, and how the development of the community will proceed (the mix of high-rises and family houses, for example, or the location of and provision of incentives for industry).

However, in the typical evaluation research design, in which a small set of goals is created by the researcher, with no input from members of the sub-units being evaluated, and then applied to all sub-units in an identical fashion, a complete homogeneity of goals within a system is assumed. In the educational system this means, for example, that all schools implementing a program or subject to a policy being evaluated are assumed to have identical goals with identical priority weightings. At the very least this will mean that schools which are pursuing different goals will not receive information relevant to their needs. At the worst, if evaluation results are taken seriously, it can contribute to an increasing homogenization of an educational system, as each school strives to increase its performance on the few selected goal measures. This not only runs counter to the generally positive valuation placed upon heterogeneity (at least in the rhetoric of public policy) but can interfere with the functioning of schools as educational institutions. As we have argued elsewhere (Alexander & Farrell, 1975):

> The zeal for uniformity has... imposed a high degree of organizational rigidity on schools and boards, which in turn has inhibited much of the creative potential of principals and teachers, as well as directors of education and trustees. It has also made experimentation with new forms of schools and board organization extremely difficult. (p. 123)

A NEW APPROACH TO GOAL SPECIFICATION

In this section we describe and discuss a new method we have developed and applied for specifying goals which may overcome the problems discussed above. At least the design was consciously constructed to avoid these problems.

The goal specification design was developed for one component of a large-scale investigation of student participation in decision-making in

secondary schools. In this component it was intended to evaluate the effectiveness of student governments in a random sample of 37 secondary schools in the province of Ontario.[4] There are two specific objectives for specifying the goals of student government (or, as we have come to call them, the "effectiveness criteria"): (a) to provide a measure of student government effectiveness which could be used as a variable for testing a series of hypotheses, and (b) to provide evaluation information to the principals of each of the 37 participating schools which would be maximally useful to them — which would identify problem areas in a manner which would stimulate them to take corrective action.

A typical design approach might have started with a search for official policy statements or statements of principals and teachers regarding the goals of student government, after which we would have tried to develop measures which more or less mapped to the goals statements, and then selected some sub-set of them for actual testing. However, asking educational officials to state the goals of their student governments evoked those typical uninterpretable and unmeasurable responses, such as "to promote good citizenship" or "to teach the value of democratic processes."

The approach we used started with the assumption that, although asking for statements of goals does not usually produce useful responses, principals do nonetheless make observations and collect evidence, even if informally, with respect to the effectiveness of their student governments — as with all other aspects of their schools. Our first task, then, was to determine what kinds of empirical evidence practising principals actually use to judge whether a student government is effective. We did this by asking a number of principals of schools not in our sample: "If you were to enter a secondary school and wanted to determine the effectiveness of the student government, what questions would you ask and to whom would you ask them?" It is important to note that we asked not for general goal statements, but rather for specific questions which could be presented to identifiable individuals, for empirical indicators of goal fulfilment. We assumed that such specific questions, reflecting evaluation criteria, could be mapped to general goals if one wished to do so. However, we believed that phrasing the question as we did would produce more meaningful responses. This turned out to be the case.

This exercise generated a long list of potential questions to be addressed to specific subgroups in a school. To this list we added a few questions which had been used in other studies, plus some which we felt, from our experience and intuitions about schooling, might be important to some principals. After checking for redundancy, a list of 49 items was finally compiled, 16 with students as the relevant opinion group, 26 with student government members as the opinion group, and 7 to be asked of teachers.

These items were then stated in an agree/disagree format and included in questionnaires to be administered to a sample of students and teachers, and to all student government members in each of the schools in the

sample. Each principal of a sample school was presented with the list of 49 items, as part of a questionnaire, and was asked first to indicate for every item how important he thought it was as a criterion for assessing the effectiveness of student government in a school like his own (Note 1). Principals were also given the opportunity to list additional criteria they thought important, but very few did so.

We next asked principals to select the six items they believed to be most important for assessing the effectiveness of their own student governments. To motivate the principals to take the selection task seriously, we indicated that because of financial constraints we could provide data regarding responses from their schools only for the six items they chose. Thus, the choice of criteria was based upon the information the principals most wanted to receive regarding their own student governments.

To reduce the possibility of principals choosing the items they believed would reflect most favorably on them and their schools we stressed that the results for a particular school would be communicated only to the principal. The principals could of course share the information with whomever they wished — and several principals later informed us that they did discuss the results with their student government members or their staff — but the choice was the principal's, not ours.

In this design we did several things we hoped would overcome resistance to negative findings. Principals chose their own assessment criteria; thus, they could not claim that the criteria used were irrelevant to their particular situations. By inviting principals to choose six items, we tried to avoid the danger of skewing the efforts of a multi-purpose organization in only one or two directions (an examination of individual principals' responses indicates that in many cases they did try to sample items which related to widely different aspects of student government functioning). By promising to communicate the results only to the principal, we tried to avoid the danger of having items chosen which would publicly make the school "look good."

Another part of the problem of resistance to negative findings has been noted by Peter Rossi (1966):

> The main reason why negative results have so little impact on organizations lies in the fact that the practitioners, first of all (and sometimes the researcher), never seriously entertain the possibility that results would come out negative or insignificant. Without commitment to the bet, one or both of the gamblers usually welch. (p. 129)

To reduce this tendency to "welch," each principal was also asked, prior to receiving any data regarding his criteria in his school, to specify for each criterion what we called the "critical decision range." This was defined as the set of values a measure must assume in order for the decision-maker to feel that a problem exists. For example, the most commonly selected item among the 49 was the following:

Members of the student government should agree that they can almost always get important changes in the school by working together constructively with the principal and teachers.

If principals who chose this item discovered that 100% of their student government members agreed that this situation existed in their school, they would clearly conclude that they had no problem in this area. If 90%, 80%, or even 70% agreed that this was the case, the principals might also conclude that they had no problem. But at some point — 60% agreement, 50%, 20%, or whatever — a principal will say: "Wait a minute. If only that many members of student government agree that they can get changes by working constructively with us, then I've got a problem on my hands." All points that produce this "wait a minute" response constitute the critical decision range. Thus, principals were asked to specify for each of their six chosen criteria the level of agreement which would indicate to them that they had a problem.

There is also the possibility that a principal, on receiving a negative finding, could rationalize it away by saying, "I knew all along that I had a problem here." To make this explicit, principals were also asked to estimate the responses of members of the relevant opinion groups in their schools for each item they chose. By comparing this estimate with the principal's critical decision range for the item, we could determine if the principal was predicting the existence of a problem.

In summary, the technique just described was designed to permit key actors in the system being evaluated (in this case principals) to participate actively in the process of specifying assessment criteria. Not only did principals assist in defining the original set of 49 items, but all chose the items they felt most important for their own schools (thus avoiding the problem of imposing homogeneity upon a complex system). Other features were built into the process to increase the probability that principals would view the results as important to them and, if the findings were by their own definition negative, to act upon them. When results were given to principals, we provided them with reminders of their original estimates and critical decision ranges, the actual responses from their school, and comparative data from the other 36 schools in the sample.

Two important questions remain: (a) Would the principals be able and willing to participate in such an exercise? and (b) Would they actually find the results of this technique useful? Regarding the first question, we can note that of the 37 principals whose schools were involved in the study, 28 were willing to participate in the entire exercise, following through on all requests for information. We have no evidence from either examination of response patterns from all the principals or from conversations with several of them that they had any difficulty understanding what it was we were asking them to do (although it is possible that those who did not participate opted out because they could not understand the procedure). Thus, we conclude that the substantial majority of principals

were both willing and able to join in this type of co-operative specification of assessment criteria.

We have not yet been able to conduct a systematic follow-up which would permit us to assess whether the participating principals found the exercise useful or acted upon any negative findings presented to them. But we have had conversations with several principals which suggest that this approach is seen as more useful than the standard evaluation research style. We know of several schools where the principal has used the results to initiate discussions with staff and students regarding the problems identified, and the principal of one school has requested a great deal of additional information on his school in order that he and his staff and students can better understand what is happening in their school. While these indications are far from conclusive, they do point in the right direction.

SOME EMPIRICAL RESULTS

Our arguments concerning the inadequacy of the typical evaluation research design are based upon several assumptions:

1. Actors in various sub-units of a system will choose different assessment criteria for their own units.

2. The assessment criteria chosen by key actors in a system will not be identical to those created by researchers.

3. Key actors in a system know a good deal about their system, are able to predict responses fairly accurately, and can identify many problems in advance of evaluation research. They are not, however, perfect predictors.

If assumptions 1 and 2 did not hold, there would be no need for the kind of approach we have developed. If assumption 3 were completely wrong, and key actors know little about their system, our technique would not work. Conversely, if actors were completely accurate in predicting responses and problems, there would be no need for formal evaluation research. The responses from the 28 principals who fully participated in the goal specification exercise permit us to test these assumptions.

What Criteria do Principals Choose?

We can test both assumptions 1 and 2 by considering the patterns of choice by principals among the 49 potential assessment criteria. While some criteria or items are more popular than others, no general consensus emerges centring on one or a very few items. The most popular item was selected by 12 (40%) of the 28 principals. One other item was chosen by 11, and 5 were chosen by 8 principals. To have invented and imposed on all schools a single universal set of assessment criteria, and to have provided information on these, would not have served the interests of (i.e., would not have provided relevant information to) these principals.

There is one other extant study which has attempted to assess the effectiveness of student governments (McPartland, et al., 1971). In that study of a number of student governments in Baltimore schools the traditional procedure of applying a single set of criteria to all schools was followed. We can test the second assumption by comparing the items developed in the McPartland study with the choices of these 28 principals.

The first McPartland criterion, "Student government can change school rules even if the teachers or principal are against the change," was not included in our list of potential criteria. We did not believe that any principal would select such an item. The second McPartland criterion was, "Student government does not worry only about social activities in the school." Three of our items related to this issue. The first was considered important by no principals; the last two were each chosen by only three principals.

The third McPartland criterion was, "The teachers and principal do let students who really disagree with them get elected to student government." Our item 39 is identical except for some minor changes in wording. Two of the 28 principals considered this to be important.

The fourth McPartland criterion was, "Students who are interested in making changes will go to the student government." While we did not present this criterion to principals directly, three of our items are close to it in meaning. These were chosen by five, one, and five principals respectively.

The final McPartland criterion was, "The people elected to student government do not simply say what will make the teachers and principals happy." This item (again with changes in wording) was included on our list. While this is one of the seven most popular items it was nonetheless chosen by only 8 of the 28 principals. Thus, even though this is the most popular of the four McPartland items directly or indirectly included in our list of assessment criteria, it was chosen by fewer than one-third of the principals.

In the preceding paragraphs we have presented evidence in support of the following points: (a) There is no single set of assessment criteria we could have chosen which would have been seen as relevant by all, or even the majority of, the principals included in our study. Indeed, there is no single item chosen by as many as 50% of the principals. (b) The one list of universal criteria developed by researchers in a previous study would not, as a set, have appeared relevant to evaluating their student governments for any of the principals participating in our study. We conclude, then, that the first two of our assumptions hold.

How Good Are Principals as Predictors?

By comparing the principals' estimations of the responses from their own schools to the actual responses, we can assess how good principals were at

estimating. There were 168 possible estimates (28 principals and six items chosen by each one). In four cases a principal did not provide estimates for all six possible items, leaving a total of 164 actual estimates. About 26% of the total number of estimates were accurate within nine percentage points and 56.7% were accurate within 20 percentage points. This means that 43.3% of all estimates were incorrect by more than 20 percentage points. It appears that principals are indeed fairly, but far from perfectly, accurate in their predictions of responses.

By comparing actual responses to both the principals' critical decision ranges and to their estimates of responses we can determine whether the principals are able to predict accurately the existence of problems in their schools. Not all principals provided us with critical decision ranges for all items chosen. However, 157 items did have critical decision ranges provided, and of these 63 (41%) did turn out to be problems.

Principals predicted that problems would exist in only 27% (43) of the criterion items. However, not all predictions were accurate; often a principal underestimated the level of agreement with an item, and thus predicted the existence of a problem where none in fact existed. Of the 43 predicted problems, 17 actually were not problems. Of the 63 problems which actually did exist, 37 were not predicted. Thus, from either standpoint — predicting problems which did not in fact exist, or failing to predict problems which did turn out to be present — these results provided a large number of surprises to principals. From all of the evidence presented above we conclude that our third assumption holds. Principals are moderately good predictors regarding their schools, but there is still much room for evaluation research to provide them with surprises.

A Note on "Goal Free" Evaluation

A basic assumption of our argument has been that an evaluation research exercise must start with identification of the goals to be evaluated. Scriven (1972) has recently introduced a fundamentally opposed notion, that of goal free evaluation. In Scriven's model, the evaluator consciously refuses to become aware of the intended goals of a program or policy, and bases judgments on whatever he/she observes to be occurring. The principal rationale for this approach is that knowing the goals in advance tends to blind the evaluation researcher to unintended consequences which may be very important outcomes. Worthen (1974, p. 21) provides an example. A program may be found to be achieving its intended goal of bringing dropouts back into school, but have the unintended result of increasing the crime rate among non-dropouts. Conversely, a program might not be achieving its intended goal (e.g., increasing reading competence) but might be producing very important results in an unanticipated area (e.g., increasing students' feelings of self-worth). In both cases, an

exclusive focus on the intended goals could produce an evaluation judgment which would lead to socially unfortunate consequences.

We accept the importance of looking for unintended side effects in an evaluation exercise. However, what Scriven (1973) has called the ideology of goal-free evaluation reflects an epistemological point of view which we do not accept. In Scriven's picturesque metaphor:

> The goal free evaluator (GFE) is a hunter out alone, and goes over the ground very carefully looking for signs of any kind of game, setting speculative snares when in doubt. The goal-based evaluator, given a map that, supposedly, shows the main game trails, finds it hard to work quite so hard in the rest of the jungle. (p. 327)

We would maintain that no evaluator enters the field without some sort of conceptual map which signals what to look for and what to ignore. For all of us, our perceptions of any given social situation are the products of our past, and are filtered through the conceptual structures already in our minds.[5] (Indeed, this is what makes the creation of goals by a single evaluator particularly pernicious.) Continuing the metaphor, we would prefer that the cartography be done by the main actors in the system to be evaluated, rather than by an external evaluator, and that the map be explicit rather than implicit.

Nonetheless, the detection of side effects can be a very important part of many evaluations. One of the advantages of the approach we have proposed is that it does not base an evaluation upon the implicit perceptual maps of any single individual (evaluator or key policy-maker). Rather, it involves many actors, and compels them to make their maps as explicit as possible. The more actors involved in the goal identification process, the greater the probability that consequences not intended by any particular policy-maker will be taken into account.

Finally, we note, as Worthen (1974) points out, "Even Scriven agrees that the major share of evaluation resources should go to goal-directed formative and summative evaluations" (p. 22). We conclude then that it is appropriate to work toward the development of evaluation techniques which will make goals the explicit public property of many actors rather than the implicit creations of individual evaluators.

CONCLUSION

In the preceding pages we have discussed a number of concerns we have regarding the way in which evaluation research in education is generally conducted. We have presented a novel technique for evaluation goal specification which, we suggest, avoids these problems, and some of the empirical results from the first application of the technique. The technique we have devised, being new, undoubtedly needs to be tested in a wide variety of evaluation research circumstances to determine its general applicability. All we can say at the moment is that it appears to have

worked well in one specific application. The technique is somewhat more time-consuming than the standard approach to evaluation goal creation, but our experience suggests that the marginal cost of doing this sort of exercise within a large evaluation study is quite modest.

One limitation of this application is that we involved only one set of actors, school principals, in the goal specification exercise. By the logic of our earlier argument we could well have involved other actors — teachers, students, parents. Part of the reason for dealing only with principals was logistic. As we were experimenting with the operationalization of this evaluation approach, time did not permit us to extend it further (deadlines are a reality in most evaluations). Another part of the reason was conceptual: We believe that, whether one likes it or not, school principals are ultimately responsible for what occurs in their schools, including their student governments. We would like to see a system develop wherein principals (and generally individuals with organizational authority) are required to "place their bets" and are held responsible for the outcomes. Under such circumstances, wise principals would ordinarily want to consult with other actors, but the choice, and the responsibility for it, would be theirs. We consider it important to avoid the troublesome consequences of diffused authority (if no one can specify who is responsible for a decision, no one can be held responsible for its consequences). However, for evaluation researchers who do not share this view, and for programs or policies where authority is necessarily diffuse (e.g., course selection in a "free option" curriculum) one would want to extend this application to include other actors. This would also be a useful test of the extent of applicability of the evaluation technique described here.

One aspect of this first application, which is both a problem and a suggestion of a long-term benefit to be derived from its wide use, relates to the inexperience of school principals with this type of exercise. To our knowledge, this was the first time anyone had ever asked these education officials to specify evaluation criteria, to estimate how people in their schools would respond to particular questions, and, most importantly, to establish a priori critical decision ranges. While most principals carried through with the exercise, understood what we were asking them to do, and produced predictions which were not wildly inaccurate, it was nonetheless a difficult and foreign task. It appears that the most difficult part of the work was the setting of critical ranges.

In regard to educational indicators (and indeed to most social indicators) we are currently where we were with many economic indicators some years ago. If one had asked a series of government officials 100 years ago to specify an unacceptable level of unemployment in their jurisdiction, they would have had the same difficulty the principals in our study had. The concept was not well defined, data were not systematically collected, and there was no general consensus concerning acceptable and unacceptable levels. Today, however, most individuals are familiar with

the term, and we have had considerable experience with unemployment data. There is a fair degree of consensus, at least in North American societies, that it is almost impossible to get the rates below 3 or 4%, that rates in the range of 6 to 8% represent economic difficulties, and that rates above that are very dangerous. We have, then, a familiar indicator with a well-defined critical decision range. (Recently, however, the federal government has been trying to persuade the Canadian people to raise their critical decision range, suggesting that the society can tolerate higher levels of unemployment.) These conditions pertain for many economic indicators. They pertain for no indicators of educational performance. We suggest that one important side benefit of inviting education officials and others within or interested in educational systems to participate in the kind of exercise we have described in these pages could well be to hasten the process of familiarizing individuals with what indicators are and how they behave under various circumstances. As with most things in life, the more people do it, the better they are likely to become at it. And the better people become at goal specification, the more likely we are to have enlightened rather than confusing public debate over those internal efficiency questions which are so salient to educators and the public today, such as, "Are the schools providing children with a good basic education?"

NOTES

[1] For an interesting example of this vagueness, see the booklet, *Policy Decisions of Board,* produced by the Toronto Board of Education. In this 150-page document, which lists all the policy decisions taken by that board in a four-year period, there is only one policy that even uses the word "goal." In 1972 the board adopted "a goal of providing for all of the educational needs of all the people in the City of Toronto" (Board of Education for the City of Toronto, 1974, p. 88). University and college presidents, adult educators, as well as taxpayers, can only hope that the board will not attempt to achieve that goal! Note also that the Report of the Hall-Dennis Commission (*Living and Learning*), charged specifically with identifying the "aims and objectives" of education in Ontario, contained the following observation in the opening section: "This Report has been designed to communicate the Committee's viewpoints, findings and recommendations in a manner which reflects the philosophy of the Committee. It contains a commentary on the aims of education, but it does not include a formal statement of aims" (Provincial Committee on Aims and Objectives of Education in the Schools of Ontario, 1968, p. 3).

[2] Simon (1964) has pointed out that "in the classical economic theory of the firm, where no distinction is made between an organization and a single entrepreneur, the organization's goal — the goal of the firm — is simply identical with the goal of the real or hypothetical entrepreneur" (p. 2). However, when dealing with educational systems (indeed with public social service delivery systems generally), one cannot assume this easy identity between organizational goals and the goals of any single individual. As Eide (1969, p. 79) points out, this is particularly a problem in educational systems. Because the ultimate "product" — actual or potential changes in the behavior capabilities of students — is the result of a long chain of interventions, means and ends are often confused. One actor's goals are another actor's means to some other (often subsequent) goal.

[3] Carter (1971) has dealt extensively with this problem in evaluation research, as has Campbell (1974). However, the observation is hardly new. In 1620 Bacon (1630/1956) said, "The human understanding is no dry light, but receives an

infusion from the will and affections, whence proceed 'sciences as one would'! For what a man had rather were true he more readily believes. Therefore he rejects difficult things from impatience of research; sober things because they narrow hope ... things not commonly believed, out of deference to the opinion of the vulgar" (p. 18). In relation to an extreme form of "bad news," Kubler-Ross notes that the first stage of reaction to news of impending death is "denial" (Kubler-Ross, 1973).

[4] Since this study involved a random sample of Ontario schools, one could legitimately question whether the principals of the sampled schools were the clients of the evaluation, in the sense of the argument just advanced. However, the study was commissioned by the provincial ministry of education because of widespread concern over then increasing student unrest in Ontario's secondary schools, which was expressing itself in various forms of student protest activities — e.g., strikes, sit-ins, demonstrations, vandalism. Concern over this problem was widespread, as demonstrated in a preliminary survey to which more than 80% of the secondary school principals in the province responded (Alexander & Farrell, 1973). In that survey, most principals perceived student protest as a significant and growing problem, and considered student government reforms an essential element in coping with the problem, and many were experimenting, or contemplating experimentation, with their student government. Thus, the principals were, albeit indirectly, involved clients. This may help account for the willingness of most of them to continue to participate in what turned out to be a very complex and, for them, time-consuming exercise.

[5] Doby (1969) has put the issue this way: "If one reacted to his visual environment with purely sense-data responses he would be judged to be out of his mind. The mind and the eye are related and the relationship between 'seeing' and the stock of knowledge one possesses is complex" (p. 139).

REFERENCE NOTE

[1] The 49 items and the format in which they were presented to principals may be obtained by contacting the authors directly.

REFERENCES

Alexander, W. E., & Farrell, J. P. Protests grow for student participation in high schools. *School Progress*, 1973, *42*, 22-24.

Alexander, W. E., & Farrell, J. P. *Student participation in decision-making.* Toronto: The Ontario Institute for Studies in Education, 1975.

Bacon, F. [Novum organum] (D. Bronstein, trans.). *Basic problems in philosophy* (2nd ed.). Englewood Cliffs, N.J.: Prentice Hall, 1956. (Originally published, 1620.)

The Board of Education for the City of Toronto. *Policy decisions of board 1969 to 1973, inclusive.* Toronto: Board of Education, Administrative Services Department, 1974.

Buck, V. The organization as a system of constraints. In J. D. Thompson (Ed.), *Approaches to organization design.* Pittsburgh: The University of Pittsburgh Press, 1966.

Campbell, D. T. Assessing the impact of planned social change. Paper presented at the OECD seminar on Social Research and Public Policies, Dartmouth, N.S., September 12-15, 1974.

Carter, R. K. Clients' resistance to negative findings and the latent conservative function of evaluation studies. *The American Sociologist*, 1971, *6*, 118-124.

Doby, J. T. Logic and levels of scientific explanation. In E. F. Borganta (Ed.), *Sociological methodology 1969.* San Francisco: Jossy-Bass, 1969.

Eide, K. The planning process. In S. Elam & G. Swanson (Eds.), *Educational planning in the United States.* Itasca, Illinois: Peacock, 1969.

Etzioni, A. *Modern organizations.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Gross, B. What are your organization's objectives? A general systems approach to planning. *Human Organization,* 1965, *18,* 195-216.

Hall, F. *Organizational goal determination in public schools.* Unpublished doctoral dissertation, University of Toronto, 1975.

Homans, G. C. *Social behavior: Its elementary forms.* New York: Harcourt, Brace and World, 1961.

Kubler-Ross, E. *On death and dying.* New York: Macmillan, 1973.

McPartland, J., et al. *Student participation in high school decisions: A study of 14 urban high schools.* Washington, D.C.: U.S. Department of Health, Education and Welfare, 1971.

Mohr, L. B. The concept of organizational goal. *American Political Science Review,* 1973, *67,* 470-481.

Perrow, C. The analysis of goals in complex organizations. *American Sociological Review,* 1961, *26,* 854-866.

Provincial Committee on Aims and Objectives of Education in the Schools of Ontario, Report of the Committee. *Living and learning.* Toronto: Ontario Department of Education, 1968.

Rossi, P. Boobytraps and pitfalls in the evaluation of social action programs. In *Proceedings of the Social Statistics Section of the Annual Meeting of the American Statistical Association.* Washington, D.C.: American Statistical Association, 1966.

Scriven, M. Pros and cons about goal-free evaluation. *Evaluation Comment,* 1972, *3,* 1-4.

Scriven, M. Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process.* Berkeley: McCutchan, 1973.

Simon, H. On the concept of organizational goal. *Administrative Science Quarterly,* 1964, *9,* 1-22.

Stewart, G. On looking before leaping. In *Social indicators: Proceedings of a seminar.* Ottawa: Canadian Council on Social Development, 1972.

Suchman, E. H. *Evaluation research.* New York: Russell Sage Foundation, 1967.

Worthen, B. R. *A look at the mosaic of educational evaluation and accountability.* Portland, Oregon: Northwest Regional Educational Laboratory, Paper Series, No. 3, 1974.

William E. Alexander is an Associate Professor in the Department of Adult Education and Joseph P. Farrell is an Associate Professor and Chairman of the Department of Educational Planning, both at the Ontario Institute for Studies in Education, Toronto, Ontario M5S 1V6.