

From Hubris to Reality:

Evaluating Innovative Programs in Public Institutions

Gerald Halpern

Gerald Halpern, Halpern & Associates

**Halpern previously worked in the
Office of the Auditor General of Canada as a
Director, Results Measurement**

Published February 1998

From Hubris To Reality: Evaluating Innovative Programs in Public Institutions

Gerald Halpern

Introduction

Le véritable manque, c'est le manque d'information qui nous font réfléchir et qui nous font créer. --Bruno Lussato

There once was a time when federal government analysts acted as if the tools of program evaluation and economic research could solve the dilemmas that surround the very difficult program choices that constantly face government. In Canada, that time was the tenure of Douglas Hartle at the Treasury Board Secretariat (TBS). It began in the late 60s and led to the establishment of a Planning Branch in 1970 which actually conducted high level evaluation studies. The United States, during the 1960s, had served as a model with the introduction of cost-benefit analysis and PPB (Program, Planning and Budgeting) techniques at the Department of Defence at the behest of Robert McNamara.

Significant shifts away from program evaluation and towards performance measurement have recently taken place in both Canada and the United States. The Canadian event that perhaps best marks the shift from arrogance to reality for evaluation practitioners would be the 1993 removal of a separate Office of the Comptroller General (OCG) of Canada, established in 1978, and the placing of its responsibilities with the Government Review Division of the Administrative Policy Branch of TBS. At the time of this writing, it rests within the Performance Management, Review and Reporting sector. Additional background reading is available in a Senate report and in a recent Auditor General report.

The parallel "end of an era" occurred in the United States in 1996 with the termination of the Program Evaluation and Methodology Division (PEMD) of the U.S. General Accounting Office. Overlapping the PEMD closing is the introduction of performance measurement requirements. The key legislations on this topic are the 1990 Chief Financial Officers Act and the Government Performance and Results Act, 1993. The former requires agencies to clearly define their mission, measure efficiency and effectiveness, and improve performance where deficient. The latter, partly as a result of increasing pressure for accountability, made performance measurement mandatory. It insists that agencies go beyond the measurement of only inputs and direct outputs. Agencies must have five year strategic plans (mission and long-term goals), annual performance plans (short term goals linked to longer term objectives) and annual program performance reports (what was actually accomplished in relation to what was intended). An insightful summary of the current situation in the United States has been given by Chris Wye in a 1994 speech: "while our political system has done its job in developing performance measurement legislation, our program managers have barely begun to take their earliest and most tentative steps toward implementation

The working definition that I will use for evaluation has two components. First, there is the discovery of a reliable difference between an expectation and an observation. Second, there is the placement of a value judgement on that variance. The emphasis in this paper, as stated in its title, is on the evaluation of innovative programs. For purposes of this paper, the following definition is accepted: *Innovation is defined as the first time a new way of doing something, a service, program or administrative technique, an approach or a technology is used in a country.*

Evaluation - General

Too bad that all the people who know how to run the country are busy driving taxicabs or cutting hair. --George Burns

In our daily lives, we evaluate. Similarly, in government we evaluate and we are evaluated. We all have a stake in government programs and we all believe that we can evaluate the utility, even the cost-effectiveness, of those programs. Since everyone outside of government knows, and knows without effort, what programs are needed and how to design and implement those programs, the people in government are either going to either have to bring in the barbers and taxi drivers as consultants or they will themselves have to engage in systematic measurement over the long haul in order to have the information needed for program improvement on a continuing basis. Not only does it make defensive sense to evaluate one's program, in the Canadian government it is incumbent upon managers at all levels to do so.

The *Review Policy, Internal Audit and Evaluation* of the Treasury Board of Canada mandates a full range of review activities ranging from "ongoing performance monitoring and self-assessment by front-line managers" to the evaluation of departmental and cross-departmental programs in support of government decision taking. The two objectives of the Treasury Board Review Policy are to ensure that the government (1) "has timely, relevant and evidence based information on the performance of its policies, programs and operations, including the results they achieve" and (2) "uses this information to improve the management and cost-effectiveness of policies, programs and operations, and to account for results".

After several careful readings of the Review Policy and in conjunction with observation of, and discussion with, line managers and evaluation managers, this writer has reached the conclusions that: (a) both of the TB objectives are incontestably good; (b) the Review Policy is biased toward evaluation that serves senior management and Parliament rather than operational line managers ; and (c) that performance measurement is both undervalued and not understood to be potentially of great value to the policy objectives.

Innovative Programs

*As the births of living creatures at
first are ill-shapen, so are all
innovations, which are the birth of
time. --Francis Bacon*

Innovative government programming is not an oxymoron. Governments, the same as private individuals and major corporations, are constantly faced with changing circumstances where old

solutions no longer work well enough. When the variance between the current solution (i.e. current program) and the intended results is sufficiently large, there is dissatisfaction ample enough to warrant the investment needed to design a new solution (i.e. a replacement program). This is the case both where a new need is recognised for which there is no current program or the current program is no longer satisfying.

Innovative programs differ from mature programs both because the program need is more salient (meaning that more people pay more attention to the need and to the proposed solution) and because the program design is more a design than a finished product. Innovative programs have to be grown from initial "ill-shapen birth" into a well formed program. The transformation typically requires three stages.

The three stages are development, installation and maintenance. The engineering costs of these three are quite different with the installation costs being very noticeably the highest. For the vast majority of government programs, the development stage occurs as a policy exercise. It may have been preceded by the results of initial "brainstorming" from a political office. At this stage, the need is recognised and the solution direction formulated. The major elements of what will be a program are more or less well mapped out at this stage. At the development stage, the theory and mechanisms of the program may be as fully explicated as is possible without the benefit of a trial run under operational conditions. This is most likely to be the case with programs of financial management (such as tax credits intended to achieve a defined goal). At the other extreme are programs with objectives and mechanisms that are much less rigorous and which are given to field staff for implementation. This is more likely to be the case for social intervention programs, particularly those which can be funded within an already existing program. These are then given to operational staff for implementation and (often unrecognised) modification. Financial programs are less amenable to field motivated modification than are interventions designed to change behaviour.

The point to be stressed is that innovative programs require field development. The amount of field development is much more extensive than is generally recognised. The bulk of the capital investment (both dollars and intellectual capital) are spent in the installation phase. Research with the French immersion programs in the schools of Ontario demonstrated that anew.. An even lengthier example is found on research with "workfare" programs.

The following chronology is taken from the twenty year history of the [United States] Manpower Research Demonstration Corporation, an agency created to provide policy makers and practitioners with reliable information on what is, and is not, effective in the design of work-focused welfare reform strategies. The intention was to derive information from the field testing of new ideas in real world circumstances and to inform decision makers on what works, and what does not work. The actual history is much too extensive for this brief article. What is instructive for those who want to understand how to develop effective innovative programs (with the emphasis on effective) can be gleaned from a brief overview of the stages of a twenty year installation of programs.

Stage One began in 1974 and largely consisted in recognising a dual problem. There were poorly defined yet real social problems to be solved and there was a need for rigorous testing of the

effectiveness of the various solution methods. The early trials taught again the following lessons. Poorly defined problems add to the difficulty of understanding the implications of studies of the results of interventions. Study findings are difficult to interpret no matter how well designed technically the studies may have been for the results will be coming from "a black box". Programs poorly documented in terms of the changes to be effected and/or the specific mechanisms by which the changes are to be brought about are very difficult to evaluate. The results of the test procedures merely result in vigorous (often acrimonious) argument among interested stakeholders, including the program designers, over the meaning of the findings. Innovative programs are purposeful attempts to try something new. Did an innovation work? We have to know in full detail just what was the purpose to be achieved (social problem to be alleviated) and we have to know exactly what was the innovative treatment received (not simply what was planned as the treatment) before we can have reliable answers to the questions.

Stage One rolled quickly into Stage Two where the focus for about five years was on testing social policies in a laboratory environment. Some ten communities with supported work programs agreed to participate in random assignment controlled experiments. "The result was the first definitive study of an employment program and the first convincing evidence that work programs make a positive difference for welfare recipients."

Stage Three took the studies out of the laboratory and into the field. It began in 1981 and continued until 1996. This was a period during which United States federal funds for state programs to change welfare into workfare was contingent upon federal approval. The result was a number of carefully designed studies. As of 1996, the federal approach removed the research component for the Aid to Families with Dependent Children program and replaced it with block grants.

Although there are many lessons to learn from this and similar evaluations of innovative programs, only four will be highlighted: make a correct diagnosis of the problem; have a reasonable treatment; find out whether the program works; and why. It costs time, effort and money to formally evaluate. It takes courage to accept that evaluation must be equally open to finding either failure or success. Careful, formal evaluation however does have an additional benefit: it yields a greater understanding of why some programs work and thereby advances the capacity of government to render its multiple missions.

Moderating the expectations held for innovative programs is the following logical anticipation. A higher proportion of innovative programs will be found wanting than is the proportion for mature programs. It is simply not reasonable to believe that all innovative programs will work. Mature programs have stood the test of time; new programs have not.

Program Theory

Alice: Would you tell me, please, which way I ought to go from here?

Cat: That depends a great deal on where you want to get to.

--Lewis Carroll

If you don't know where you are and as well do not know where you want to be, no amount of information can make your journey more efficient. In the absence of pre-set goals, there is no point to evaluation. Exhibit 1 presents a generic program model designed to illustrate the theory of a program and the chain of results which the program intends to produce.

The key elements of this exhibit recognise that there is a program which receives inputs (resources) which it transforms into outputs. The leverage areas for program management are program plans, resultant program organisation, implementation of that organisation, and internal control mechanisms for keeping the program on track. Legitimising any government program is a legislative authority defining the desired outcomes and providing resource supply. Senior executives and policy planners elaborate the program and its objectives and "hand" it to operational managers to make it work. The effectiveness of the program is judged in terms of its chain of results: outputs (some of which are products and services internally servicing the program operations); impacts that flow from the outputs; and effects upon society. The impacts and effects are jointly be referred to as outcomes. Program results include both outputs and outcomes.

Not shown in this diagram is the full panoply of environmental factors that influence all aspects of the program beginning with the form and level of legislative direction, continuing with the details of the program formulation and concluding with the pushes and pulls that interact with the program for the shape of actual results.

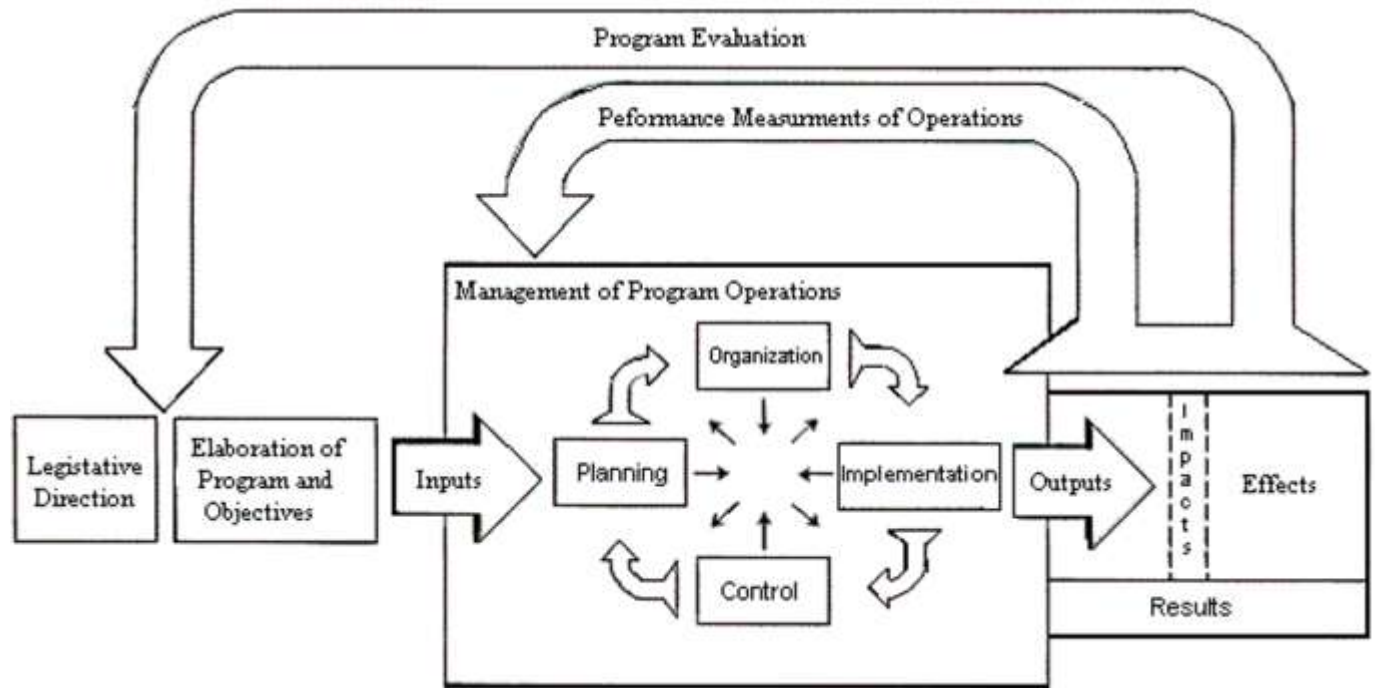
Evaluation is a control mechanism that enters the picture in a variety of ways. It is a process used by organisations to measure the performances (results) that flow from designated operations. It involves agreeing on clear goals, developing performance indicators to track progress, establishing baseline data, setting targets for future performance and periodically gathering actual results data for comparison with the targets. The two broad evaluation perspectives are program evaluation and performance measurement.

For program evaluation, the key issues are whether the observed results may reasonably be attributed to the fact of the program and whether the observed results of the program are judged to be of value. The focus will be on program outcomes, those longer term results which are expected to satisfy the program's objectives. This is the effectiveness issue and it is here that program evaluation places its focus. Outcomes information is intended to inform program continuation decisions.

Performance measurement also looks at results but tends to focus on the early results, the more direct outputs of the program. Attribution at this level is often taken for granted. The focus on outputs serves the management concern for program improvement. Key elements of this concern are effectiveness (Are targeted outputs being met?) and efficiency (Has the ratio of outputs to resource inputs been maximised?). Performance information is designed to inform resourcing and program design decisions.

The theory of a program is the set of imputed links between program operations and program results (outputs, impacts and effects). Strong theory (theory with evidence supported links) increases the probability that the desired outcomes will occur and that the program will be the most plausible explanation for the observation of the outcomes.

Figure1: Exhibit 1: Schematic Representation of a Program



Program Realities

Nothing is impossible if you can only assign it to someone else to accomplish.

The reality is that evaluation, on the one hand, is conceptually easy and, on the other hand, is very difficult in practice. Conducting trustworthy evaluation is not easy. It is especially difficult in the public sector. The two most important factors that make government programs difficult to evaluate (and of course difficult to manage) are goal complexity and weaknesses in the means-and-ends linkages of a program's theory.

Goal Complexity The ability to measure performance is inextricably linked to a clear understanding of what a program is trying to accomplish. If you don't know what you are looking for, how will you know when you have found it? Worse, what will prevent you from claiming that whatever you do find is just what you were looking for?

Education is a form of government programming. As an example of lack of goal clarity, let us consider recent occurrences in high school mathematics grading. A mathematics teacher has in his class a student who is bored, a mischief-maker, speaks a non-official language at home, and who has a very poor command of English. This student scores 100 on the mathematics examination. According to a provincial department of education, that person's score in mathematics should be reduced from 100 by deducting marks for misbehaviour and for poor English. Is the mathematics course objective to be "proficiency with the knowledges and skills of the mathematics curriculum" or is it "mathematics performance plus proficiency in the English

language plus classroom behaviour"? If the latter, then justice demands that the set of objectives, plus their relative weights, be announced in advance.

In government, it is typical to find that there are multiple goals and they may be contradictory. The several goals are often in competition with each other and their salience at any given future moment is difficult to predict for their relative importance is in constant flux. Public sector programs typically have results intentions in three areas: administrative (with the goal of efficiency), political (with the goal of responsiveness to constituencies demands), legal (the legal rights of individuals must be fully respected). All are important in their own right.

All three co-exist in all government programs. They compete with each other for managerial attention. It is only reasonable to expect that resources for any one objective will be less focused and accomplishments reduced given the multiple objectives that draw upon the resources of the program. Add to this the general reluctance to assign additional resources specifically to the task of bringing a program design to mature program status. The close observer of actual government operations might be forgiven for having some admiration for the performance of individual program managers.

A recognition of goal complexity changes the way in which performance is measured and evaluated. The performance of government managers must be assessed against the full range of goals. To merely focus on efficiency and then to make comparisons between government program managers and other managers focused only on an efficiency (profitability) objective would be an egregious case of invidious comparison.

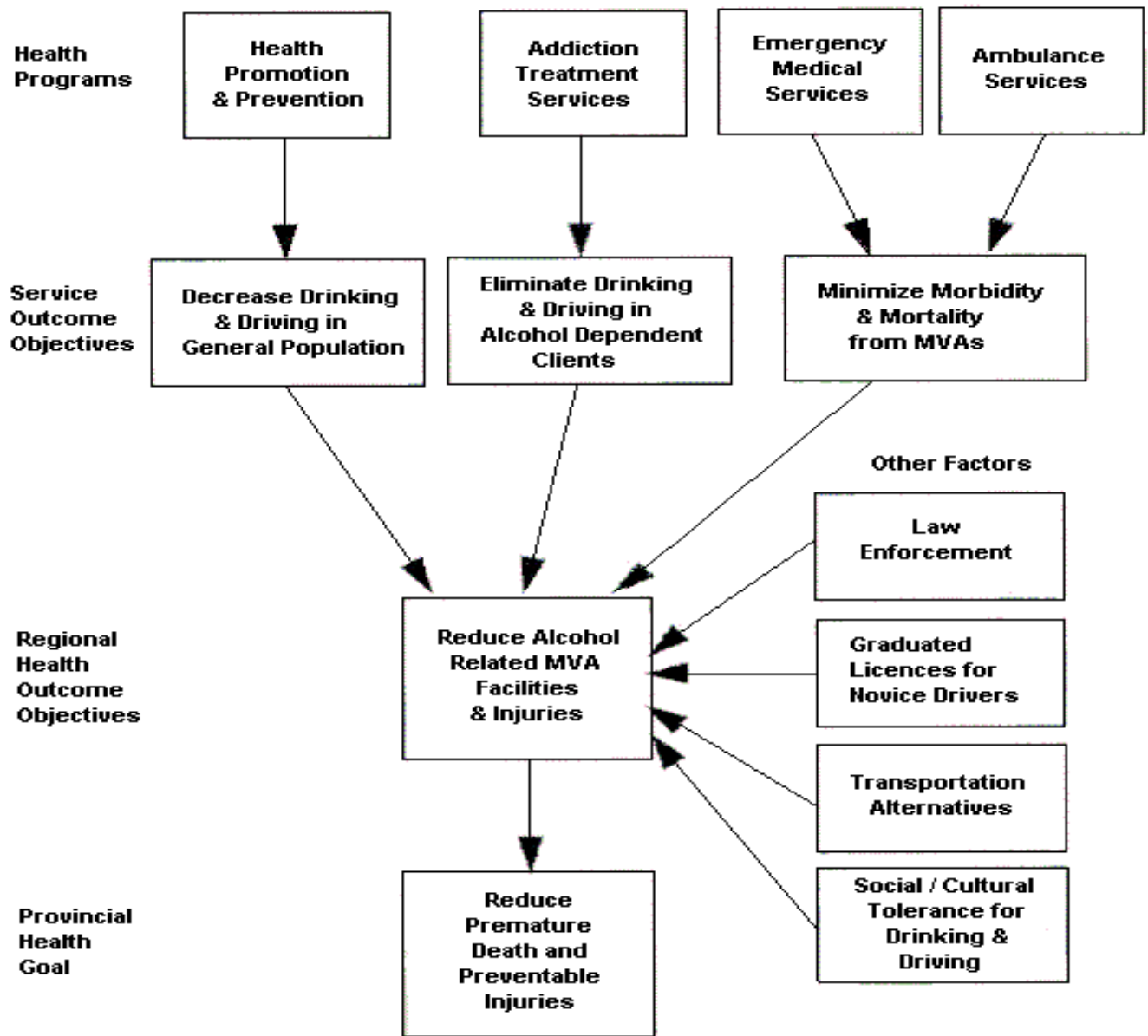
Means and Ends The "means" designed for a program are intended to lead to the intended effects. A program requires a sustaining theory. There has to be good reason to believe that because we do A, B will occur which in turn will yield C.

The following example illustrates a proposed causal linkage for a government program in British Columbia. The model is reproduced in full in Exhibit 2. This form of a logic model is conceptually the same as that shown in Exhibit 1. It differs in its de-emphasise of the presentation of program operation and its focus on program results including specific inclusion of factors other than the program that may influence the intended program outcomes.

The authors argue that the development of this program logic model (program theory) "facilitates overall governance of health care services by creating performance-monitoring frameworks for both short term and long-term outcome objectives."

Is Exhibit 2 representative of strong theory? That is an empirical question deserving of proper reply. It deserves reply for it is the path to program improvement. In order to evaluate program impacts, we must have a theory of why the intended impacts will occur. In order to improve program operations, we must have a theory of how operations produce outputs. Innovative programs tend to have less well supported causal chain linkages than mature programs that have been moulded by years of experience. Performance information will help the program's operations to more efficiently produce the desired outputs. During this formative period, it is also good strategy to seek evidence that the outputs lead to the impacts and effects in intended ways. In time, program evaluation will test the mature program against its outcomes.

Figure1: Exhibit 2: Logic Model for Reducing Alcohol-Related Motor Vehicle Accident Injuries and Deaths



Without such theory, evaluation findings are mere facts. They lack the context to turn them into information capable of program improvement. Unfortunately, many public programs lack strong theory. Good research requires time, money and senior level attention. Few would argue that either innovative or mature government programs are blessed with the apparent luxury of good research. This paper argues that true program understanding is a necessary condition. Innovative programs, which always begin as first time efforts, in particular need evaluation derived knowledge to serve as performance feedback for the guidance of program evolution.

The Evolution of Innovative Programs

The Road to Wisdom? Well, it's plain and simple to express:

Err and err and err again but less and less and less.

Piet Hein, Grooms, 1966

The message of this paper is straightforward. When an innovative program is the object of evaluation, the role of evaluation is that of servant to the program's design. Evaluators should work with innovative program managers to strengthen the program theory, to test operational elements of the program, to formalise the trials by which the program is made increasingly efficient in producing outputs and to conduct the necessary research to test the linkages between operations - outputs - impacts - effects. These roles together are the means by which evaluation can serve the objective of program improvement and control.

Returning to Exhibit 1, evaluation in service to innovation would work with the program manager inside the box labelled "Management of Program Operations". Primary attention would be given to the management functions of planning and control. Evaluation information for planning would focus on program theory testing, both within the program operations and equally important, the testing of the sequential linkages of the results chain. There are evaluation tools available ranging from evaluation synthesis to direct empirical studies. Evaluation serves the management control function by the identification of key variables, development of measures for them, appropriate measurement activities, analysis and interpretation, and presentation of findings on a schedule to match management decision cycles. Working with a program in development (which always describes the early stages of any innovative program), the evaluator works for the program staff with the intention of improving the program. Questions of effectiveness in producing outputs are primary followed by efficiency monitoring as the program staff searches for ways to develop and improve the program.

Using evaluation to produce better programs is never a substitute for program evaluation as a tool for evaluating a mature program in terms of its effectiveness or its program efficiency. Referring again to Exhibit 1, the key question for program evaluation is whether the program results are consonant with the program intentions. For government funded programs, the source of the program's intentions must be the legislative direction that gave rise to the program objectives. The innovative program in government is an attempt to satisfy a new legislative direction. Once the program reaches maturity and has been there for a period of time sufficient to attain the objectives, the program is ready for program evaluation.

Unfortunately, the Review Policy of the Government of Canada does not encourage the use of evaluation resources in support of innovative programs. Neither is such use prohibited for the policy does not prevent evaluation as a tool for operational managers. There is mention of performance measurement in the review section of the policy and that section recognises that managers would benefit from consultation with program evaluators. But the policy does not encourage the role of performance measurement. The policy provides an annex for Program Evaluation and another annex for Internal Audit. Performance measurement does not receive parallel attention. Nor does the policy speak to the use of evaluation for anything other than mature programs. The special needs of innovative programs are not recognised in any way. This

is a review policy that stresses the utility of evaluation for senior managers and Parliamentarians. Given the context of scarce resources (managerial think time being among the scarcest), it is not surprising that ongoing performance measurement is little used in a formal sense.

Coda

One caveat is in order. This paper has argued for a greatly enhanced role for performance measurement within the range of evaluation approaches favoured for Canadian government managers. It is not likely to happen unless it is valued by the departmental career gatekeepers. It has to be rewarded in its own right and it must be integrated with program evaluation if it is to be believed. Performance measurement requires staff conviction and staff empowerment. Giving people the capacity and authority to evaluate their programs on an ongoing basis is not enough. Busy people set priorities. They do what is important to them and what is important to their superiors in the administrative hierarchy. Empowerment has to be within an organisational culture that encourages and rewards continuous learning. Staff will learn more about their programs and will use the learning to improve program if they recognise that to be valued by their corporate culture. From my experience, I must add that the people best placed to improve programs are those who work directly with the programs and who have a ego investment in making them into quality programs. They are the organisational force for continuous improvement rather than major re-engineering efforts.

About the Author: Gerald Halpern, Halpern & Associates

Dr. Halpern has worked in a number of government departments. Before opening his own consulting boutique, he was employed with the Office of the Auditor General of Canada as a Director, Results Measurement and Audit. Dr. Halpern has just completed his term as an Ottawa Board of Education School Trustee.