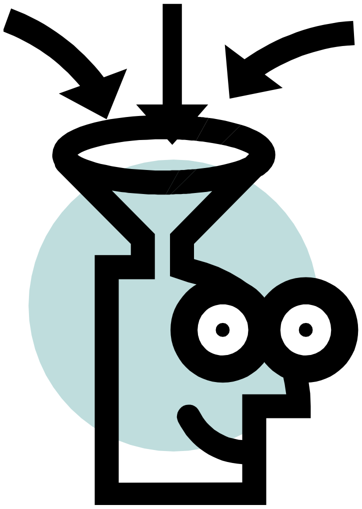
Mark Schacter
CONSULTING

Tell Me What I Need to Know

A practical guide to program evaluation for public servants



by Mark Schacter

schacterconsulting.com

Tell Me What I Need to Know
A practical guide to program evaluation for public servants

Published by:

Mark Schacter Consulting
Ottawa, Ontario, Canada

www.schacterconsulting.com
mark@schacterconsulting.com

Table of Contents

Foreword

A news item headlined ***Corporate tax cuts don't spur growth*** reported that tax cuts provided to Canadian corporations were not having their intended impact on corporate spending and job creation.[1]

The article said that the government had claimed that "tax cuts are crucial to stimulate job creation and make Canada more competitive on the global stage." But an analysis of official statistics showed that corporate "investment in machinery and equipment has declined in lockstep with falling corporate tax rates over the past decade." Instead of putting the excess funds created by tax cuts into job-creating investments, firms had "added $83-billion to their cash reserves since the onset of the recession in 2008."

The headline would have you believe that the government's corporate tax policy was a mistake because it did not support the intended outcome of corporate investment leading to job creation. This is a plausible conclusion, but is it correct?

It doesn't require an advanced degree in economics to see that the headline writer may have failed to consider other possibilities. Is it possible, for example, that the drop in corporate investment in machinery and equipment would have been *even larger* if the government had not cut taxes? If so, then we might conclude that tax cuts did in fact stimulate corporate spending even though the overall spending trend was negative.

The one indisputable fact is that the outcome of immediate interest – corporate investment in machinery and equipment – has many causes. Tax rates may affect it, but so too will the broader economic and political climate. Shifts in interest and exchange rates, stock and commodity prices and international relations will all affect the willingness of corporate executives to invest rather than hoard excess cash. It is conceivable that corporate tax cuts might induce corporations to invest, but that the combined negative impact of other factors will be greater, leading to a net decline in investment. Before pronouncing with certainty on the effectiveness of tax cuts as a stimulus of corporate investment, we would therefore have to account for the impact of other factors.

Even if we were able to establish with a reasonable degree of confidence that tax cuts had stimulated corporate investment, which in turn had stimulated job creation, we might want to ask a further question. If the ultimate goal was job creation, would it have been possible to get an equivalent or better result by an intervention other than tax cuts that would have been no more costly, or perhaps less costly, to the government?

This way of thinking reveals an "evaluation mindset". It's an approach that identifies missing information, challenges facile assumptions, and seeks gaps in reasoning. It accepts nothing at face value, but instead asks, "What is really going on here?" In the public sector an evaluation

[1] *Globe and Mail,* April 6, 2011.

mindset is, above all, about identifying value to the public – in this case, job-creation – and assessing whether or not public programs and policies are generating it in the best possible way.

In fact, more often than not it is very difficult to figure out (i) what public value looks like; (ii) whether a given public program has generated public value; and (iii) whether some other type of intervention would have done a better job of generating public value. Coming up with robust answers is an evaluator's job. It demands intelligent inquiry and subtle analysis rather than – as in the case of the newspaper story – a hasty leap to conclusions.

If you believe that the public sector has an important role to play in improving our lives, then it follows that government decision-makers need reliable evidence about how best to create public value through programs and policies. When done well, evaluation is an important source of this evidence.

Tell Me What I Need to Know
A practical guide to program evaluation for public servants

If the things we learn do not have much practical value,
perhaps we are investigating questions that are not important.[2]

A. Introduction

Elected officials define a vision of the public interest (that is what election platforms are about). They couch the vision in broad terms, painting a picture of a happier future in shades of prosperity, health, security, environmental sustainability, etc.

Public servants implement the politicians' vision through policies and programs that transform broadly stated goals into real benefits eventually felt by the public. Public service is a practical vocation: it's about getting things done against a backdrop of vaguely stated objectives, volatile circumstances, incomplete information, insufficient time and constant public scrutiny.

The title of this paper acknowledges the connection between evaluation and the practical imperatives facing public servants. First, if programs and policies are supposed to deliver benefits to people, then what do public servants *need to know* to ensure that programs and policies are designed and implemented in the best possible way for delivering benefits? An evaluation report has no value unless it provides operationally-relevant analysis and recommendations supporting decisions that contribute to better program/policy design and implementation.

Second, what should public servants be able to extract from evaluation to answer citizens' demands for accountability? The Government of Canada's accountability regime emphasizes results-based management and results-based reporting to Parliament.[3] Public managers are required to produce evaluations assessing the contribution of programs to social or economic results ("outcomes") that touch the lives of citizens.[4] They must be able to use evaluation to tell a credible story about how public interventions have made life better for Canadians.

Third, how much do public servants need to know about the inner workings of evaluation itself? Many public servants may be involved in the preparation of, or be users of, evaluation results, but few of them are or need to be evaluation experts. Instead, they need to be familiar with key evaluation concepts, and with the strengths and limitations of evaluation. They need to

[2] *Sources of Power. How People Make Decisions,* by Gary Klein, Cambridge: The MIT Press, p. 6.
[3] Canada is not alone in this. The move toward "results-based" approaches to public management is common to many other countries.
[4] The Government of Canada's *Policy on Evaluation* is found at www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15024&section=text . The related *Directive on the Evaluation Function* is at www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15681

understand how to *use* evaluation to help them be better managers. They need to know enough to be able to pose the right questions to – and make sense of answers from – evaluation experts who design and implement evaluation studies.

A typical member of this paper's target audience[5] is likely to be involved in some way in the preparation of an evaluation, perhaps as the person with overall responsibility for the evaluation. Such a person will rely on others for technical expertise; he might direct the work of evaluation consultants or be the point of contact with an in-house evaluation unit. Even when not directly involved in preparing evaluations, the people for whom this paper is intended may read evaluation reports to glean lessons relevant to the design of new programs, or because they are looking for ways to make an existing program better.

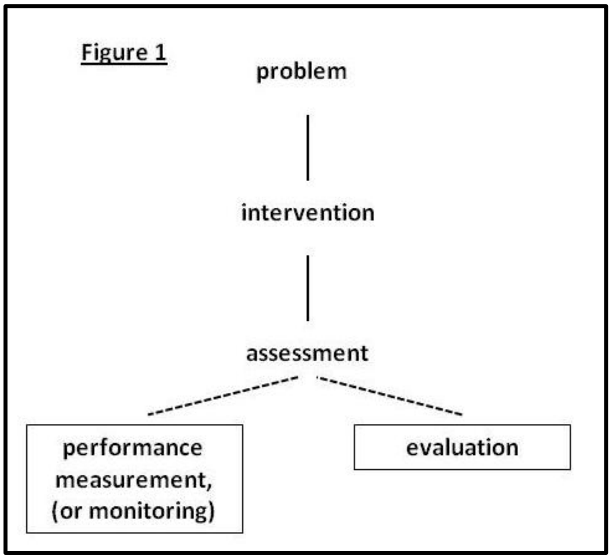
In short, this paper does not target technicians. It does not delve into technical questions. Readers looking for detailed instruction on research methods will not find it here. My aim is to give non-technical readers enough knowledge about evaluation to help them recognize how evaluation can be used to meet their needs, and to understand in general terms the technical questions that evaluation experts may bring to their attention.

B. "Evaluation" and "Performance Measurement"

The details of an evaluation vary from one project to the next, but the underlying context never changes. There are three elements (Figure 1):

- an issue of public concern – a **problem** – that the government has chosen to address;
- an **intervention** through which the government addresses the issue; and
- an **assessment** of the intervention's performance in relation to the problem.

Figure 1



For example, to help unemployed people (*problem*) the government operates labor market programs (*intervention*) that provide income support and job-training. The government routinely examines these programs to see if it they are achieving their expected results (*assessment*). Evaluation is one "assessment" piece in this pattern, but not the only one. "Performance measurement" is another assessment tool – one that is sometimes confused

[5] The primary audience for this paper is Canadian federal public servants. The paper therefore contains references of specific interest to them. However most of the content is relevant, I believe, to a broader audience.

with evaluation. The confusion is to some extent a matter of vocabulary. The term "performance *measurement*" can be taken to suggest that there is one, and only one, program management instrument that deals with performance. But evaluation, too, is a form of performance measurement. Using the term "performance *monitoring*" instead of "performance measurement" (I use the terms interchangeably) might help address the problem by making clearer the distinction between performance measurement and evaluation.

Consider an example outside the world of public management. My doctor listens to my heart, takes my blood pressure and reviews blood tests. These are *monitoring* data that provide evidence about the performance of systems in my body at a moment in time (the moment when the tests were done). They are a basis for speculating on what has happened to me in the past, and what may happen in the future. If my blood pressure is high, the doctor may suspect that I do not get enough exercise or have a diet too high in salt. She may tell me I have a high risk of stroke and heart disease.

After the appointment the doctor orders a detailed work-up: more tests, scans, and details about my medical history and lifestyle. With the new information the doctor prepares an in-depth assessment – an *evaluation* – that attempts to explain *why* my blood pressure was high and recommends steps for addressing the problem.

The distinction between monitoring and evaluation is thus two-fold:

- description versus analysis; and
- real-time, "in-the-moment" performance information versus a deeper assessment of performance over a longer period of time.

The doctor's initial reading of my data gave her an *indication* that the performance of my body's systems might be below acceptable standards. More in-depth information was required to produce a definitive diagnosis of my condition and recommend corrective action.

Monitoring tells a story about *what is happening now*. It is done continuously and provides updated performance information over a relatively short time cycle. But every once in a while it is necessary to make a more thorough assessment of performance – one that covers multiple lines of evidence, uses data representing years of performance (rather than single moments in time) and goes beyond explaining what is happening now to an analysis of *why* things are as they are, and how things might turn out differently. This is evaluation. It is a time-consuming and relatively costly undertaking, but may yield a level of understanding that cannot be obtained through monitoring.

Consider labor market programs again. Canada's Department of Human Resources and Social Development routinely collects data on, for example, the percentage of unemployed Canadians eligible to receive employment insurance payments who actually collect payments, and the proportion of Canadians who return to work or school after participating in a skills-development program. This is monitoring data; it is routinely and frequently collected, and gives program managers an idea of how things are going at the moment.

These data are necessary but not sufficient for developing a well-rounded understanding of program performance. From time to time managers of labor market programs may want to do evaluations seeking answers to deeper questions such as, *Are employment insurance payments set at the correct level? (Should we be paying more? less?) Are the rules for qualifying for program benefits too loose/tight? Is job-training provided through the program meeting the needs of unemployed people? Is the program well managed? (Are applications processed efficiently and accurately? Are benefit cheques sent out on time? Are training programs delivered properly? etc.)*

Evaluation goes beyond issues of day-to-day performance. It may look into fundamental questions of program objectives, design and management. While typical monitoring questions can almost always be answered by simply *counting* something – e.g. *the percentage of program participants who . . .*; *the number of beneficiaries who . . .* – evaluation questions usually demand a deeper level of inquiry and a more sophisticated level of analysis.

Monitoring and evaluation are complementary. Each can help managers make decisions and report on program performance. Each has strengths and limitations. Monitoring gives managers a constant stream of information required for day-to-day operations and is *relatively* inexpensive to implement, but is not suited to supporting in-depth understanding of factors driving program performance. Evaluation, on the other hand, helps managers go beyond description of the here-and-now to understand whether programs are doing what they are supposed to be doing, and why they are successful (or not). In comparison to monitoring it is costly, time-consuming and disruptive to program implementation. As a consequence, it cannot be undertaken very often.

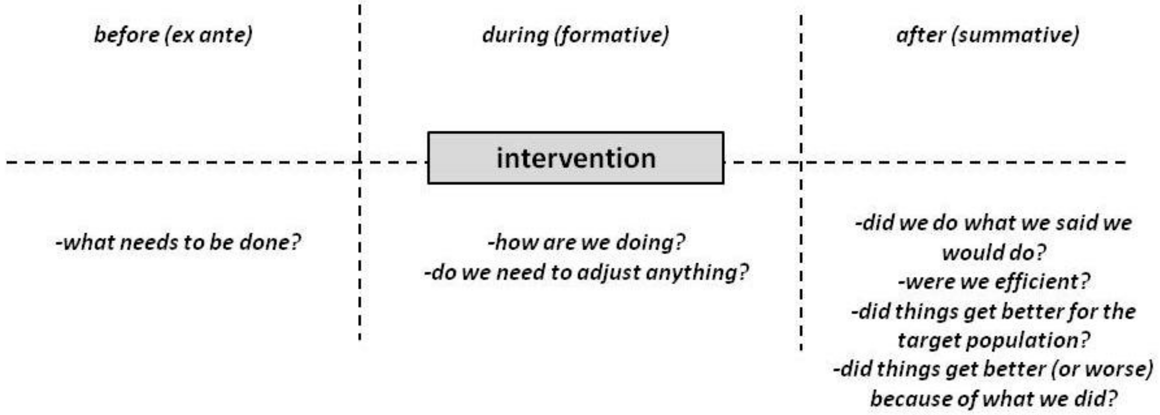
C. The Evaluation Continuum

Evaluation responds to managers' needs for information about programs; information requirements will change according to circumstances. Evaluation can be done before a program has been launched, during implementation, or, in cases where a program has an end-date, after the program is over. An evaluation can pose questions about what should be done, how things are going, or what has been accomplished.

The range of possible evaluations occurs on a continuum (Figure 2). When designing a new program, it is normal to do a study that defines and analyzes the problem to be addressed by a program, identifies target beneficiaries and defines program outputs appropriate to the nature of the problem and the characteristics of the beneficiaries. This is sometimes called a "needs assessment" or an "*ex ante*[6] evaluation".

[6] A Latin term meaning "before the event".

Figure 2


before (ex ante)
during (formative)
after (summative)
intervention
-what needs to be done?
-how are we doing?
-do we need to adjust anything?
-did we do what we said we would do?
-were we efficient?
-did things get better for the target population?
-did things get better (or worse) because of what we did?

Once an intervention has been launched and has operated long enough for meaningful trends to be established, managers may undertake what is usually referred to as a “formative” or “implementation” or “process” evaluation focused on determining whether the program is running as well as possible, and assessing the likelihood that outputs will be delivered as planned and outcomes achieved. The primary purpose is to identify changes that may be required to the design or management of the program due to unanticipated factors that have emerged during implementation.

Questions typically asked during a formative evaluation include:

- is the program on track to deliver its intended social or economic outcomes?
- is the program on track to deliver its intended quantity of outputs? is the delivery process efficient? is it effective, i.e. reaching the intended beneficiaries?
- are beneficiaries sufficiently aware of the program? are they having difficulties gaining access to it? do adjustments need to be made to the target beneficiary population?
- are internal administrative processes efficient and effective?

Finally, evaluations may be done after a program has ended (or is approaching the end of its life), or, in the case of programs that do not have a definite end date, after the program has been in operation for a significant length of time. These evaluations are typically referred to as “summative” or *ex post*[7]. As their name suggests, they are meant to provide information that allows program managers, as well as external stakeholders to “sum up” the performance of a program.

[7] A Latin term meaning “after the event”.

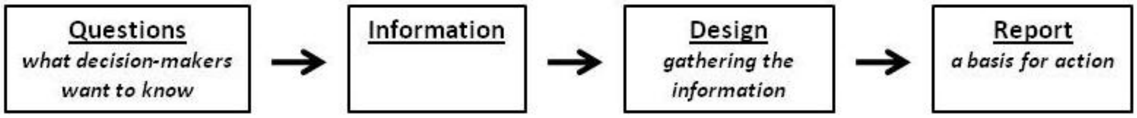
Summative evaluations, like formative ones, ask questions about program design and administration, delivery of outputs and achievement of outcomes. But a formative evaluation usually focuses on identifying corrective action to be taken during implementation, while a summative evaluation seeks firm conclusions about program performance and lessons to guide future programs. A summative evaluation may be used to decide whether a program with a pre-defined lifespan should be closed or extended. For a program that is ongoing, summative evaluations may point to a need for significant changes in program management or design – e.g. new ways to deliver the program, changes in the mix of outputs, adjustments to the target population or to intended outcomes.

A special type of summative evaluation – *impact evaluation* – is singled out for extra attention in this paper (Section E) because of its importance and the unique challenges it creates for evaluators. Impact evaluations are supposed to determine if programs have contributed to improving social and economic conditions in Canada. They play a central role in public accountability and performance reporting. If well done, impact evaluations allow Departments and Agencies to describe to Parliament, in a credible manner, what was achieved for Canadians through public expenditure. By the same token they can be used by Parliament to hold the government accountable for the way it has spent public funds.

D. Core Considerations

An evaluation – even one where the data-gathering is thorough, the analysis impeccable and the conclusions beyond dispute – *is worth nothing in and of itself*. The worth of an evaluation hinges on the extent to which it meets the needs of decision-makers.

Figure 3



It follows that for an evaluation to have a good chance of being useful, attention must be paid to four core considerations (Figure 3):

- the **questions** to which decision-makers want answers;

- the sources of **information** required to develop credible answers to the questions;

- the **evaluation design** (the way in which the information will be gathered); and

- the **evaluation report**, which delivers the evaluation's findings and conclusions to decision-makers in a way that provides a basis for action.

(i) Evaluation Questions

If an evaluation is about coming up with answers to questions, then it's important before launching the work to agree on the kinds of questions to be asked. Disappointment in the results of an evaluation – along the lines of "I have no use for these findings and recommendations" – may lead to the conclusion that the evaluator did not perform well. While this may have been the case, it is also possible that the program manager did not communicate well to the evaluator the questions to be answered.

Typical evaluation questions address:

- **Needs.** *What needs to be done?* This is the central question in a "needs assessment" or "*ex ante* evaluation". This question is crucial in the design phase of a new program when the importance of understanding the nature of the problem to be solved, and the best means for addressing it, is paramount.

- **Effectiveness.** *Did the program deliver what it was supposed to deliver?* An assessment of effectiveness focuses on the extent to which the program's outputs – e.g. income support or job training in the case of a labor market program – were delivered as planned.

- **Efficiency.** *What did we get for what we spent? Did we minimize waste?* Efficiency is typically assessed in terms of cost per unit of output. If a program delivers job-training at a cost to the program of $100 per person/day of training, but training of comparable quality could have been delivered for $60, then the program is inefficient. Sometimes evaluators are also asked to assess the cost of achieving outcomes (as opposed to delivering outputs). So if the expected outcome of job-training is that unemployed people acquire skills they need to find work, the evaluator might be asked to assess the average cost to the program of re-integrating an unemployed person into the job market.[8]

- **Program Management.** *Was the program well run?* Evaluation of program management covers the range of internal rules, practices and procedures related to day-to-day delivery of the program. It may cover issues invisible to clients (e.g. information management, financial management, quality control, internal communications) as well as management issues at the point of contact with clients (e.g. quality of client service, accessibility of services, etc.).

- **Relevance.** *Did we do the right things? (Is the program delivering the right set of outputs?)* Relevance is about the relationship between the intervention and the problem it is meant to address. Suppose that a program for unemployed people

[8] The Government of Canada's *Policy on Evaluation* distinguishes between "efficiency" and "economy", defining efficiency as the cost of producing outputs and economy as the cost attributed to achieving outcomes.

provided training on how to maintain and repair steam engines. Because there is no market for this skill the training would not help its clients. Even if participants graduated with a high degree of skill in steam engine repair, the training would be *irrelevant* to the need identified as the basis for the program.

In some contexts[9] relevance may also mean *Is the program relevant to our mandate?*, or, *Are we the right organization to be doing this?* Relevance in this sense is assessed in terms of organizational mandate (*should another Department or Agency be doing this instead of us?*) or jurisdiction (*should another level of government be doing this instead of us?*) or assumptions about the role of the public sector (*should an organization in the private or not-for-profit sector be doing this instead of us?*).

- **Impact.** *Did the program make things better for its targeted beneficiaries?* Impact evaluation looks beyond a program's success in delivering outputs (effectiveness) to its contribution to outcomes. It asks "so what?". Training may have been delivered as planned, but so what? Did it affect reintegration of the unemployed back into the job market (which was the purpose of the program in the first place)? "Impact" is both the most important type of evaluation question, and the most difficult to address, given the information and analysis required to obtain a reliable assessment. Special challenges related to impact evaluation are discussed below.

- **Sustainability.** *Is the program likely to survive without us?* Questions about sustainability are relevant when a program aims to "kick-start" something that is supposed to take on a life of its own after the program ends. It will be important to assess whether beneficiaries appear motivated and able to seek alternative ways to sustain the flow of program benefits.

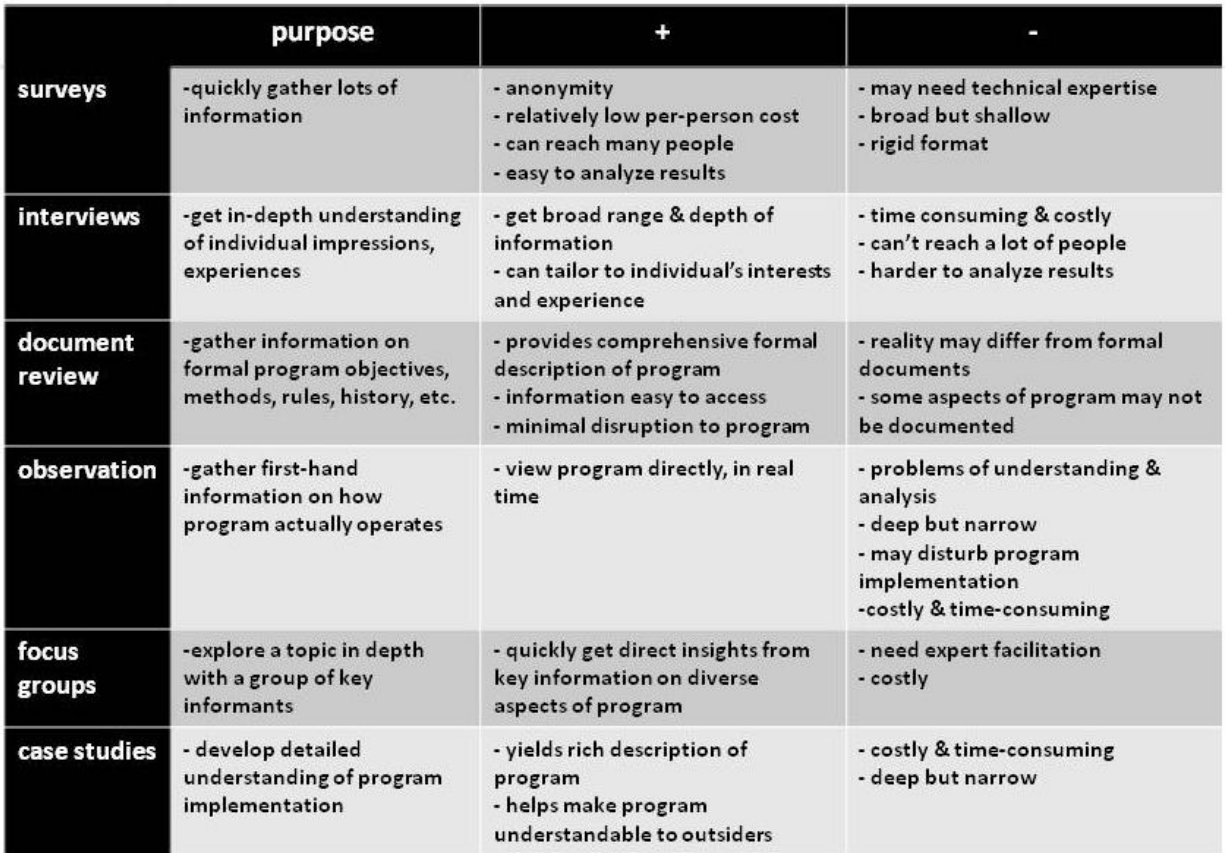
(ii) Sources of Information

Table 1[10] summarizes a typical range of information sources used in evaluations. The table points to the value of *multiple lines of evidence*. Different information sources provide different perspectives on program performance, giving the evaluator a rounded picture. As well, because every information source has its flaws, the strengths of some may compensate for the weaknesses of others. And because a critical limiting factor in any evaluation is the budget, using a mix of information sources – some relatively cheap and others relatively expensive -- helps the evaluator get maximum value from available funds. More expensive information sources such as interviews and case studies can be used in areas where in-depth information and rich personal perspectives are deemed especially important. Less expensive sources, such as standardized survey instruments or document reviews, may suffice elsewhere.

[9] As in the Government of Canada's *Policy on Evaluation*.
[10] Adapted from http://managementhelp.org/evaluatn/fnl_eval.htm

The remaining two core issues – evaluation design and the evaluation report – are discussed below in Sections E and H.

Table 1



	purpose	+	-
surveys	-quickly gather lots of information	- anonymity - relatively low per-person cost - can reach many people - easy to analyze results	- may need technical expertise - broad but shallow - rigid format
interviews	-get in-depth understanding of individual impressions, experiences	- get broad range & depth of information - can tailor to individual's interests and experience	- time consuming & costly - can't reach a lot of people - harder to analyze results
document review	-gather information on formal program objectives, methods, rules, history, etc.	- provides comprehensive formal description of program - information easy to access - minimal disruption to program	- reality may differ from formal documents - some aspects of program may not be documented
observation	-gather first-hand information on how program actually operates	- view program directly, in real time	- problems of understanding & analysis - deep but narrow - may disturb program implementation -costly & time-consuming
focus groups	-explore a topic in depth with a group of key informants	- quickly get direct insights from key information on diverse aspects of program	- need expert facilitation - costly
case studies	- develop detailed understanding of program implementation	- yields rich description of program - helps make program understandable to outsiders	- costly & time-consuming - deep but narrow

E. Evaluation Design: The Special Case of Impact Evaluation

(i) Asking an Impossible Question

Evaluating impact – the extent to which a program *caused* social or economic gains for its targeted beneficiaries – is as difficult as it is important. This is not to say that the other types of evaluation do not pose major challenges. Producing a good needs assessment depends on skillful definition of the problem to be solved and careful matching of the problem with interventions judged to be technically, fiscally, administratively, politically and jurisdictionally feasible. Implementation assessments present challenges related to, among other things, assessing efficiency – the attribution of costs to outputs and to outcomes – which is fraught with the need to make difficult assumptions.

But impact evaluation stands apart because it poses a question that is both supremely important and inherently impossible to answer with certainty. Impact evaluation poses the question we care about most, namely, *'did the program make a difference?*". As mentioned, Canadian federal Departments and Agencies are required to evaluate the impact of their programs; this requirement reflects a demand to know more than just whether or not programs are delivering what they said they would deliver, efficiently, within the bounds of law, regulation and policy. Stakeholders also want credible evidence that programs are contributing to social and economic outcomes; the attempt to provide such evidence is the defining feature of impact evaluation.

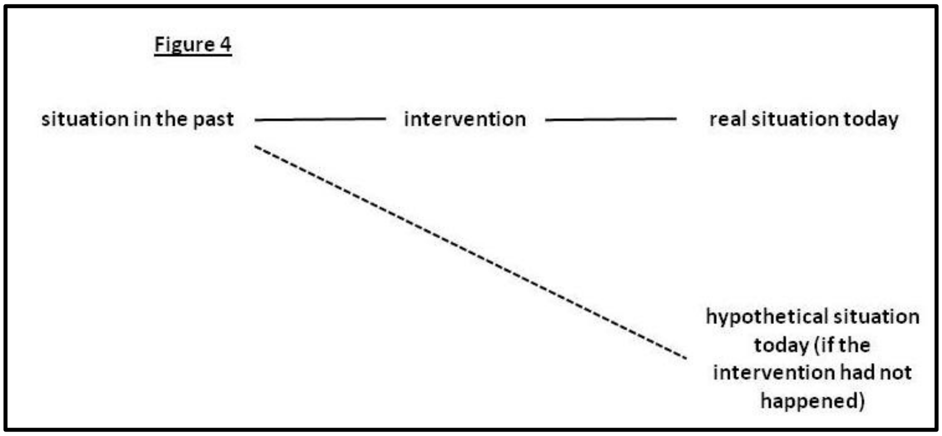


Figure 4
situation in the past
intervention
real situation today
hypothetical situation today (if the intervention had not happened)

Second, impact evaluation poses an impossible question – *what would have happened if this program had not been implemented?* This "counterfactual" question is at the core of all impact evaluation. Counterfactual thinking is inherently speculative because we can't know for certain what the present would look like if the past had been different. The quality of impact evaluation therefore depends on the quality of the evaluator's speculations. He has to produce an *approximation*, built on assumptions, models and educated guesses, of what would have been the case today if the program had never been implemented (Figure 4).

Although it may sound exotic, we all engage in counterfactual thinking. For example: "It only rains when I forget my umbrella. *If I had remembered my umbrella today, it wouldn't have rained.*" This is a common (if not not-very-rigorous) form of counterfactual thinking! Politicians will, when it suits them, pretend to have a clear picture of the counterfactual. Think about this the next time you hear a Prime Minister say, "Thanks to our policies, Canada's economy grew by leaps and bounds last year." This claim can neither be proven nor refuted with certainty. We can only speculate. A doubter might respond with a counterfactual proposition such as, *"The economy would have grown just as much, or more, if you had done nothing!"* (And we'll never know whether he or the Prime Minister is right.)

The credibility of a counterfactual is judged by the quality of the analysis that underpins it. If criticism of the Prime Minister's claim about the success of his economic policy is based only on personal distaste for the politician, then the counterfactual has no credibility. But what if it was based on analysis of national economies similar to Canada's, where policies like those of our government had *not* been implemented, and what if it found that they grew by as much as or more than the Canadian economy? In this case the counterfactual would have merit.

(ii) Creating an Alternative Reality

An impact evaluation compares the actual state of things with an alternative reality – a hypothetical world where the intervention is assumed not to have happened. The evaluator asks two key questions: (i) is the actual state of things better than, worse than or no different from the hypothetical state? and (ii) if there are differences, can they be attributed to the intervention?

The evaluator aims to construct a counterfactual that comes as close as possible to looking like the world as it would have been if the intervention had not happened. The question of "evaluation design" covers a range of factors related to how an evaluator builds the counterfactual case. There is a vast expert literature on the design of impact evaluations. The subject can be hugely complex, even controversial, and demand very high levels of specialized technical expertise.[11] This paper only scratches the surface. As stated in the Introduction, the intent is not to turn the reader into an evaluation expert but rather to provide the non-expert with a broad-brush overview of key issues.

As an entry into the challenges of designing an impact evaluation, I use a hypothetical example of a public program. Suppose a provincial government implemented a one-year pilot program targeted to people seeking their first driver's license. One out of every five first-time applicants for a driver's license was selected, at random, to complete a special accredited driver training course designed by the government. (Applicants excluded from the pilot might have chosen to take some other form of driver training, but none will have been given access to the special training program.) Two years after the pilot ended – to give participants time to accumulate a driving record – the government evaluated the pilot to see if the accredited training had an impact on road safety. How might the government have designed the evaluation, i.e. how could it have constructed the counterfactual?

a) Option 1: Experimental Approach

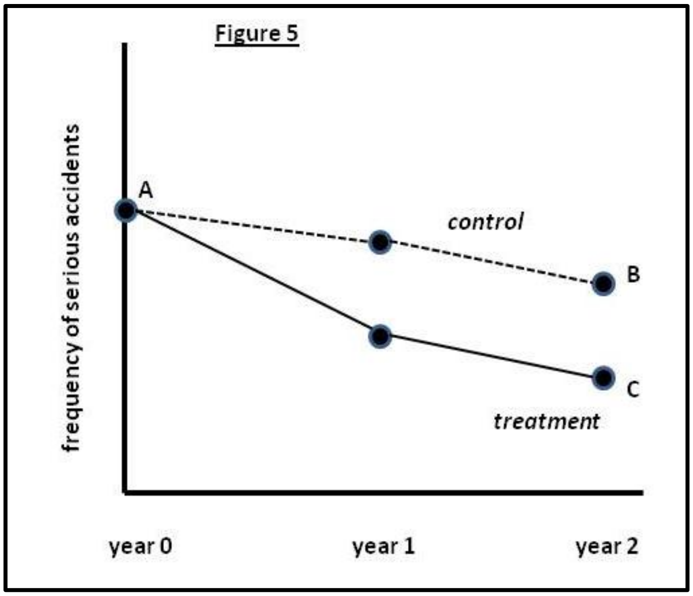
Because non-participants in the pilot were excluded at random, they are assumed to be like the other drivers in every important respect *except for non-participation in the training*. Their driving records are the counterfactual: they are assumed to represent what the driving records of the participants would have been if the pilot program had never happened. In the language of impact evaluation, non-participants are the "control group" and participants are the "treatment group". When the pilot was launched the government had data linking driving records with years of driving experience, and therefore knew the rate at which newly-licensed drivers without the special training were involved in major road accidents. This provided the year 0 "base case": how the world looked when the pilot began.

[11] White (2006) provides a useful introduction to the technical complexities and controversies surrounding impact evaluation.

The evaluation suggests that the mandatory training generated benefits because after two years the participants had better driving records than the control group (Figure 5).[12] Of course, factors other than training may have affected driver performance. Perhaps new driving laws came into effect, enforcement was enhanced, major road improvement projects were completed, or winters were unusually mild. But the power of random selection is that it automatically addresses the impact of all causal factors – both those we are aware of and those we can't observe. We can assume that all causal factors will affect randomly selected control and treatment groups equally, and that any *difference* in performance – the safer driving record (B - C) of the treatment group – is due to the only factor assumed to distinguish one group from the other, namely the special training.

Because baseline data (point A in Figure 5) were available it is possible to report not only on the final accident rate of participants vs. non-participants, but also on:

- the level of improvement shown by participants (A - C);
- the level of improvement shown by non-participants (A - B); and
- the difference between the improvement shown by participants vs. non-participants ((A - C) - (A - B)).



This is all useful information that will support operationally relevant conclusions and recommendations. But in this case, even if we had no baseline data, the endpoints alone (B and C) would show that the program generated benefits. This is because the control and treatment groups were randomly selected, which means we can assume that the starting point for both groups was identical. As described below, this is not always the case for evaluation designs that use control and treatment groups.

The evaluation design just described – the key feature of which is *random selection of control and treatment groups* – is referred to as an **experimental design.** Conventional wisdom has it that experimental design is the most robust approach to evaluation. Its key feature is that by creating a control group identical in

[12] I will not discuss the question of a "statistically significant" difference between points B and C in Figure 5. On technical points the reader is advised to consult the technical literature on research and evaluation methods.

every significant way to the treatment group it delivers results that are relatively easy to understand and come as close as possible to solving the counterfactual problem.

b) Option 2: Quasi-Experimental Approach

It is not always desirable or possible to use an experimental evaluation design. An important limitation is that it may be impractical or inadvisable to randomly select individuals to receive program benefits while randomly withholding benefits from others. Although few people might complain about being excluded from the "benefit" of mandatory driver training, it would be different if the government were offering a $100 tax credit for driver training. Randomly excluding otherwise qualified people from *that* benefit would create an uproar!

As well, there will be cases where the decision to evaluate is taken well after a program has been launched, by which point it is too late to create random treatment and control groups. Suppose that the compulsory driver training program had been applied to *all* new drivers right from the start. After the program has been running for three years an impact evaluation is launched. Although it is too late at that point to do a randomized experiment, it is still possible to have treatment and control groups. The control group would have to be *created*, rather than randomly selected, by identifying a population similar to the treatment group that had not been exposed to a training program like the one in question. If this was a provincial program, the control group would have to be found in another jurisdiction because all new drivers in the home province would have participated. A control group might be constructed from the driving records of new drivers in another province. If these drivers were found to have a worse safety record than the treatment group, there might be a basis for concluding that the program generated benefits.

The deliberate (rather than random) selection of a control group whose members are assumed to be similar to the treatment group is referred to as a **quasi-experimental design.** Quasi-experiments, if well done, can allow much of the rigor of randomized experiments in cases where randomization is either impossible or impractical. A key challenge, however, is to ensure that the method used for constructing the control group[13] produces a sample that really is similar in all important ways to the treatment group. If it isn't, then the results of the evaluation may be misleading.

Suppose that mandatory driver training pilot was introduced in Ontario in 2000. The evaluation compared the 2002 driving records of year-2000 program graduates to the 2002 records of drivers in Quebec who received their first driver's license in 2000 (the control group). Suppose the data showed that new Ontario drivers had fewer major accidents. Can the evaluator immediately conclude – as he could do in a randomized experiment – that the mandatory driver training was responsible for the difference?

[13] There is a range of options for matching treatment and control groups in quasi-experiments, some of which involve advanced statistical techniques. For an overview, see Bamberger (2006).

He cannot. Quasi-experiments are a weaker demonstration of the counterfactual than random experiments. Randomizaton, if properly done, will virtually eliminate all explanations for the outcome of interest (driving record) apart from the treatment (mandatory training) because there will be no systematic differences between the control and treatment groups. All causal factors other than mandatory training will have had the same impact on both the control and treatment groups. In a non-randomized quasi-experiment, however, we can almost never be certain of a perfect match between the control and treatment groups. The evaluator has to be alert to *alternative explanations* (apart from the treatment) for the outcome. For example, suppose that in Quebec:

- drivers are licensed at a younger age than in Ontario;
- roads are generally in poorer condition than in Ontario;
- snow tires are required by law in winter, which is not the case in Ontario; and
- speed limits are more strictly enforced than in Ontario.

The first two factors could result in new Quebec drivers having a poorer record than new Ontario drivers *even if there had not been mandatory training in Ontario*. Thus the evaluation might exaggerate the impact of mandatory driver training. (The relatively poor record of new Quebec drivers might be primarily attributable to licensing age and road conditions, rather than the training program in Ontario.) Conversely, the last two factors might improve the performance of new Quebec drivers relative to their Ontario counterparts. If these factors were dominant, the evaluation might understate the impact of mandatory driver training. (The gap between the safety record of Ontario and Quebec drivers would have been *even greater* if not for Quebec's snow-tire law and the stricter speed-limit enforcement.)

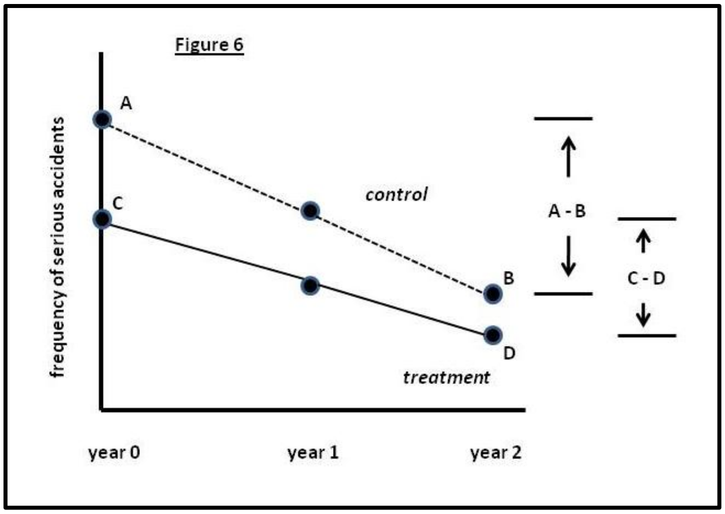
If the evaluator was able to identify key differences in factors affecting new Ontario and Quebec drivers, he might be able to account for them. He might match the age profile of drivers in the control and treatment groups; through statistical modeling he might approximate the influence of differences in road conditions, use of snow tires and strictness of speed-limit enforcement. These "biasing" factors will inevitably crop up in quasi-experiments because creating a perfect match between the control and treatment groups is difficult. An evaluator will attempt to remove as much bias as possible, but it is unlikely that he can remove it all. There may be practical constraints – for example, it takes time, money and data to build statistical models. And in many cases the evaluator may simply not be able to observe key characteristics that distinguish the control and treatment groups.

Suppose the evaluator was unaware of stricter speed limit enforcement in Quebec. Suppose he had neither the time nor the budget to construct models approximating the impact of road conditions and mandatory use of snow tires. “Double-differencing”[14] is an example of a common and relatively simple technique for controlling bias caused by “unobservable” influences. Figure 6 illustrates double differencing.

Notice that in Figure 5 (p. 12) the control and treatment groups are assumed to have the same frequency of serious accidents in year 0 – a reasonable assumption given random selection from the same eligible population. In Figure 6 the control and treatment groups are at different points in year 0 – to be expected given the difficultly of precise matching when random selection is not used.

The different starting points affect interpretation of the data. In a random experiment it is meaningful to know simply that there is a statistically significant difference between the end points (B and C in Figure 5, B and D in Figure 6). Not so in a quasi-experiment. Figure 6 shows that after two years of mandatory government-approved training, new Ontario drivers have fewer serious accidents (point D) than new Quebec drivers (point B). But if we compare the *change* in the control (Quebec) group’s performance (A – B) to the treatment group (C – D), we find that the control group had a *greater improvement* than Ontario – in other words, [(A – B) – (C – D)] is a positive number. This suggests the need for skepticism about claims that the better safety record of new Ontario drivers “proves” that the program was successful. Indeed it appears more likely that the Ontario program had no positive impact at all, and may even have been counterproductive!

Double-differencing is supposed to account for some of the bias caused by systematic differences in the control and treatment groups. The degree of change experienced by the control group is assumed to be the counterfactual – the amount of change that would have happened anyway (sometimes referred to as “natural change”), without the intervention. If the treatment group has more than the natural degree of change, then there may be grounds for concluding that the program had an impact.



An important condition for using double differencing is availability of both *before* and *after* data for the control and treatment groups. Another is that the differences between the two groups must remain constant over time. Double differencing wouldn’t work if, for example, the Quebec Provincial Police relaxed their enforcement speed limits during the period covered by the evaluation.

[14] Sometimes referred to as “difference-in-differences”.

Differing circumstances, budgets and objectives of a given evaluation study may lead to less rigorous, less reliable variants of the quasi-experimental design. One example would be *before*/*after* data for the treatment group, and only *after* data for the control group. A practical advantage of this approach is the reduced cost of data collection that results from eliminating baseline data for the control group. Another variant, frequently used, is to have only *after* data for both the treatment and control groups. The further loss of reliability that results from eliminating baseline data may be regarded as an acceptable tradeoff in some cases given the additional savings in data-gathering costs.[15]

c) Option 3: No Control Group

A Canadian public servant might question whether the ideas in the preceding section have anything in common with impact evaluation as it is actually done in the Government of Canada. Experimental and quasi-experimental designs are the exception rather than the rule the Canadian public service; it is indeed quite possible that many public servants have never been involved in or read reports based on evaluations of this type.[16]

The predominant impact evaluation design in the Canadian government (and there is no reason to believe that the situation is different in other jurisdictions) looks only at the treatment group – sometimes with but often without baseline (or "before" data) – and does not include a control group. For the driver training example, this would mean drawing conclusions on the impact of mandatory training by looking only at the "before" and "after" performance of Ontario drivers, or perhaps by looking only at their "after" performance. The absence of a control group makes this the weakest evaluation design – *the one least capable of providing reliable evidence of a program's impact*. This is because the story it claims to tell about impact ignores the influence of external forces and is based instead on the questionable assumption that observed benefits are entirely due to the intervention.

There may be sound reasons for choosing this weak evaluation design. Many of the interventions that public sector evaluators must assess are not suited to using a control group. Impact evaluation of an internal service or process (e.g. a corporate human resources group, or a corporate policy group) is a case in point. These "programs" are universally applicable to certain classes of transactions and/or are universally available to users. Finding a control group within the same organization as the treatment group is impossible, and finding a "comparable" control group in another organization is impossible in practice if not in theory.

Consider the Treasury Board Secretariat's 2009/10 evaluation of the Treasury Board submission process. (The submission process is the vehicle through which federal organizations seek

[15] Adapted from Bamberger (2006).

[16] Using publicly available information, I reviewed a non-random selection of 20 program evaluations produced by 17 federal Departments and Agencies. Of the 20, three evaluations incorporated quasi-experimental design elements, and in one of those three the design was acknowledged to be significantly flawed.

approval from the Treasury Board (a Cabinet committee) for initiatives they would not otherwise be able to undertake.) The evaluation sought to assess, among other things, the impact that this often lengthy and intricate process has on the quality of decision-making by Treasury Board Ministers. This is a case where there simply is no control group. The only even remotely imaginable possibility would be a ministerial decision-making body in another jurisdiction, but the difficulties in accounting for differences in decision-making processes, let alone gaining access to study such a group in detail, would be overwhelming. The only available option is indeed to assess the apparent impact of the intervention (Treasury Board submission process) on the treatment group (Treasury Board Ministers).

The difficulties are not limited to internally-oriented interventions. Many outwardly oriented public programs are also not well suited to a control group/treatment group approach. One example is the 2010/11 evaluation by Fisheries and Oceans Canada of the International Fisheries Conservation Program.[17] According to the evaluation report, the program

> promotes and protects the interests of Canadians by ensuring access for Canadians to fish resources managed internationally. [It] promotes and influences sustainable regional fisheries management and healthy global marine ecosystems, and contributes to a stable international trade regime for Canadian fish and seafood products. This is achieved through a coordinated and proactive approach, building broad and constructive relationships with international partners based upon common goals and strategies.

The program appears to be so broadly targeted and susceptible to a wide range of external influences that it may be difficult to identify the treatment group – let alone a possible control group – with precision![18] These general features are not uncommon to many public programs.

Even in cases where it may be possible to incorporate a control group into an impact evaluation, it may be decided that time or budget constraints make this infeasible. Whatever the reason for choosing an impact evaluation design that examines only the treatment group, it is important to recognize the limitations of this approach. To be sure, a thorough and detailed study of the treatment group – especially one that includes baseline data – may provide useful information on whether and why outcomes occurred, and the effectiveness of administrative arrangements, etc. It may support informed speculation about program impact. But if there is no control group, an impact evaluation is unlikely to provide a reliable answer to *the* critical question: what would the outcome have been for the treatment group in the absence of the intervention?

[17] www.dfo-mpo.gc.ca/ae-ve/evaluations/10-11/6b121-eng.htm

[18] Apart from posing challenges for an evaluator, this raises important fundamental questions about the program itself.

F. The Importance of Being Logical

There are many reasons why programs succeed or fail. Less than total success might be attributable to poor program design, incorrect assumptions, flawed implementation, the unanticipated impact of external factors beyond the control of the program, or, most probably, some combination of all these things. It follows that if evaluations are going to help program managers manage better, then they must not only show that an intervention worked (or did not work) as intended, but must also provide insights into *why* this was the case.

Evaluations that don't go beyond describing impact are sometimes called "black box" evaluations because the factors that contributed to impact remain inaccessible to decision-makers. A key to opening the black box is to spell out the program's underlying theory or logic. Readers familiar with performance measurement will recognize this as the "logic model" that should be at the heart of a framework for measuring (monitoring) a program's ongoing performance. A logic model identifies the key actions ("activities") undertaken in a program to produce products or services delivered to beneficiaries ("outputs"), and then links the products or services to social or economic consequences ("outcomes"). The logic model explains why a program was implemented. It says: *we decided to invest in delivering output X because we believed (for reasons A and B) that it would contribute to outcomes C and D.* Articulating a program's rationale in this way provides a basis for asking intelligent questions about why the program was or was not successful.

Figure 7

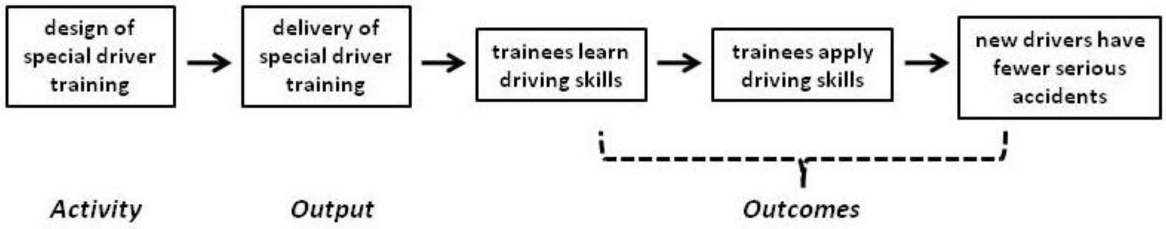

design of special driver training
delivery of special driver training
trainees learn driving skills
trainees apply driving skills
new drivers have fewer serious accidents
Activity
Output
Outcomes

Figure 7 illustrates the logic model of the mandatory driver training program. We can infer from it a rationale for the program along the following lines:

- new drivers are having an unacceptably high rate of preventable accidents;

- an important cause of the accidents is inadequate training of new drivers;

- participants in a well designed and well delivered special training course will learn skills to help them avoid serious accidents;

- new drivers will apply these skills; and

- there will be a lower rate of serious accidents among new drivers.

The logic model points the evaluator to key issues that must have been on the minds of the program's designers. It focuses his attention on questions to explore during the evaluation – questions that will help to explain the program's performance.

Suppose for example that the evaluation used the quasi-experimental design described above, whereby the treatment group in Ontario was matched with comparable new drivers in Quebec. And suppose it was found through a double-difference calculation that there was no significant difference between the performance of new drivers in Ontario and Quebec. The logic model leads the evaluator to questions that are most likely to explain the apparent ineffectiveness of the program. For example:

- Was the initial observation about new drivers in Ontario having an unacceptably high rate of serious accidents correct, or was it based on defective data?

- Was lack of skill the primary factor behind the serious accidents, or was something else responsible (e.g. road conditions)?

- Did trainees learn valuable driving skills in the mandatory training course? Were there flaws in the design or the delivery of the training course that impeded learning?

- Did trainees fail to apply the skills that they learned in the course?

- Did the training emphasize the right skills, i.e. skills linked to prevention of the types of major accidents that were actually occurring?

In an ideal case, the evaluator will find that a logic model already exists for the program under review, but this is not always the case. When a logic model cannot be found, the evaluator should try to piece one together based on information available in program documents and from people knowledgeable about the program.

G. Quantitative vs. Qualitative Methods

The professional evaluation community, like any other professional community, has its share of debates and disagreements about technical questions. An issue that generates considerable heat has to do with the place of qualitative as opposed to quantitative methods. Some argue that rigor in evaluation is synonymous with a heavily quantitative approach. According to this view, evaluations must be founded on careful statistical analysis of meticulously gathered data – preferably data comparing randomly chosen treatment and control groups – leading to conclusions stated with scientific accuracy.

The contrasting view is that evaluators often assess programs whose performance is the result of complex human behaviors and institutional and jurisdictional interactions that cannot faithfully be reduced to numbers. Context, it is argued, is at least as important as numerical data points – and acquiring an adequate understanding of the context of program design and implementation requires qualitative data gathered through observation, discussion, interaction and review of background documentation.

As you might expect, the optimum approach will probably be found somewhere in the middle. Public servants would be uncomfortable (under most circumstances) with heavily quantitative evaluations because the world of randomized controlled experiments (RCEs) seems distant from the world of public policies and programs. RCEs are suited to interventions where one (or a small number of) independent variable(s)[19] can be tested on a treatment group in isolation from other independent variables that might also affect the outcome. An example that non-experts intuitively understand is a randomized drug trial, where an experimental drug is given to a randomly selected group of qualifying subjects, while a placebo is given to a randomly selected control group. In this case there is a distinct problem to be addressed (an illness – the “dependent variable”), a single, clearly identifiable independent variable (the drug) and well-defined treatment and control groups that can be isolated from factors that may bias the experiment. The outcome of interest is easily defined and measured: the treatment group will or will not have a different experience with the disease than the control group.

Few public sector interventions fit this template. If we think back to the example of the Treasury Board submission process, the “independent variable” is not one discrete item like a drug. Rather, it is a way of doing business – a set of rules, procedures and relationships meant to affect the quality of decisions taken by senior officials. The treatment group – decision-makers – cannot be isolated from biasing factors outside the submission process, and in any case the biasing factors are difficult to specify. The dependent variable – the quality of decision-making – is itself difficult to define and measure with precision. The only feasible evaluation design is a non-experimental one (no control group), and most likely a design that uses only “after” (as opposed to before and after) data. As noted, this is the least reliable type of evaluation – it gives us the least amount of confidence in drawing conclusions about impact – but under the circumstances it is the best that can be done.

On the other hand, there are public sector interventions that do lend themselves to varying degrees of quantitative, experimental evaluation. In the driver training example the independent variable (training) is easy to identify, and the dependent variable (driver performance) is also easy to identify and measure. Biasing influences (road conditions, traffic-

[19] An independent variable is the factor that is deliberately manipulated to affect the treatment group. In the example of the driver training program, the mandatory training program is the independent variable.

law enforcement) can be identified and incorporated into the analysis. A quasi-experiment appears feasible here. The same could be said for any type of intervention where

- there is a relatively discrete output (e.g. funding, training, information);
- biasing factors are relatively limited in number and can be identified and assessed; and
- the intended outcome can be relatively clearly defined and measured (e.g. changes in economic welfare, employment status, health, awareness or capacity to perform certain functions).

In the driver-training example, a quasi-experimental design gives us more confidence than a non-experimental to draw conclusions about impact. But, as noted in the discussion about logic models, the quasi-experiment on its own has nothing to say about *why* the training was effective (or not). The "why" question is what is most important to decision-makers, and answers to it are likely to come from qualitative methods: interviews with participants and instructors, observation of training sessions, review of course materials, etc. A combination of quantitative and qualitative methods will probably be the best way to produce meaningful, operationally relevant evaluation results in cases like this one.

H. The Evaluation Report

The report is the evaluator's opportunity to pull all the pieces of the evaluation together into a story that is useful to decision-makers. Everything discussed so far – evaluation questions, information-gathering methods, evaluation design, logic models, qualitative vs. quantitative approaches – is worth nothing in the absence of a well-crafted evaluation report.

A decision-maker, not surprisingly, is looking for analysis to help him make well informed decisions. A good evaluation report will help the decision-maker think clearly about next steps to be taken in relation to the evaluated program, e.g. should it be redesigned? should it be managed differently? does its allocation of human or financial resources need to be adjusted? Or it may help him decide about the design, management or resourcing of another program that has features in common with the evaluated program. If an evaluation report doesn't help him to do these things, then his response to the report is likely to be "so what?".

Evaluation reports tend to be organized around variations of a structure where you find:

- an **introduction** providing the name of the program evaluated; the circumstances of the evaluation (e.g. an end-of-program review); the questions the evaluation was supposed to address (e.g. "The purpose of the evaluation was to assess the impact of mandatory driver training on the rate of serious motor vehicle accidents among new drivers in Ontario."); and any other general contextual information;

- a **short description of the program** covering its essential features, including intended outcomes, products/services delivered, target population, key aspects of the program's delivery mechanisms, management structure and governance;

- a non-technical **description of how the evaluation was done** ("methodology"), including evaluation design, data sources, time-segment covered by the evaluation (e.g. "The evaluation reviews the 2002 driving records of new drivers who received their licenses in 2000."), and notable limitations of the evaluation (e.g., "Due to the high rate of staff turnover in this program, the evaluator was unable to arrange interviews with a significant number of key staff.").[20]

- the **findings of the evaluation**; in other words, a summary of key determinations of fact arrived at by the evaluator;

  - e.g. "We found that new drivers in Ontario who completed the mandatory training had a rate of serious accidents that was 15% below that of drivers in the Quebec control group."; and

- a **concluding section** that gives the evaluator's analysis of what the facts mean, and makes **recommendations**; rarely will the facts speak for themselves, and so an important part of the evaluator's role is to interpret facts in a way that has meaning for decision-makers;

  - e.g. "Although Ontario drivers had a lower rate of serious accidents than the Quebec group, it was not possible within the time and resource constraints of this evaluation to rule out, conclusively, influences other than the mandatory training, such as poorer road conditions and lower age limits for driver's license eligibility in Quebec, which may also have had a significant influence on the rate of serious accidents. However, bearing all factors in mind, as well as the perspectives of key informants, our conclusion is that mandatory training is likely to have had a significant positive impact on driver performance, over and above the influence of other factors. In particular, we believe that mandatory driver training was responsible for reducing the rate of serious accidents among new drivers by somewhere between 0.5 and 1.3 accidents per 10,000 kilometers driven. If decision-makers seek greater certainty and precision regarding the impact of mandatory training on driver performance, we recommend additional research to distinguish more clearly the effect of training from the effects of the major external factors noted in the evaluation.

[20] General readers of an evaluation report are unlikely to be interested in a detailed technical description of methodological issues. A thorough technical description should however be included in an annex to the report so that technical experts can assess the quality of the approach taken.

The extent to which the evaluator adheres tightly to a particular organizing framework for the report is less important than the extent to which the report itself provides clear answers to questions that are of greatest importance to the decision-maker. As he works his way through the report, it should be easy for the reader to identify:

- this is what was evaluated;
- this is what the evaluation tried to figure out;
- this is how the evaluation was conducted;
- this is what the evaluation found;
- this is what the findings mean for decision-makers; and
- this is what decision-makers with an interest in this program should consider doing now.

As obvious as this may sound, far too many evaluation reports either do not provide all of this information and analysis, or do not make it easily accessible to the reader. In other words, too many evaluation reports are written in a way that leaves decision makers asking "so what?".

I. Conclusion

Evaluation provides in-depth, evidence-based assessments of various aspects of program performance – everything from whether a program is well-managed to whether it is well designed, whether it delivered its outputs as planned or whether it made a significant contribution to intended social or economic outcomes. This paper focused on the latter point: *impact evaluations* that attempt to assess a program's contribution to positive social or economic changes experienced by targeted beneficiaries.

Emphasis on impact mirrors the practice of evaluation in the Government of Canada. The *Directive on the Evaluation Function* requires that every evaluation examine, among other things, "progress toward expected outcomes … including the linkage and contribution of outputs to outcomes."[21] The government's focus on evaluating impact makes sense. Parliamentarians and citizens care most, at the end of the day, about whether or not public spending *makes things better* for individuals, companies and organizations targeted by public programs. This should also be the primary concern of program managers. Emphasis on impact is a cornerstone of both good management and meaningful public accountability.

What is less obvious – indeed, the point is not addressed in the official documentation – is the high level of difficulty faced by an evaluator seeking to assess the impact of a public program on

[21] Annex A of the *Directive.*

desired social or economic outcomes. Impact evaluations present daunting technical problems that other types of evaluation do not. This paper described the challenges of creating the counterfactual. It noted that experimental and quasi-experimental research techniques regarded as the most reliable methods of creating the counterfactual are often deemed impossible, very difficult, or impractical in relation to many types of public programs. The paper observed that the evaluation design normally regarded as *least valid* for assessing impact – i.e. an assessment of the treatment group only (no control group), often with only "after" as opposed to "before and after" data – is commonly used in the Canadian government for assessing the impact of public programs.

The point of the latter observation is to highlight the importance of approaching evaluation – whether you are involved in designing and implementing an evaluation, or are simply a reader of an evaluation report – with an informed mind and a critical eye. *You should be aware not only of the strengths of evaluation, but also of the important limitations of evaluation, as it is currently practiced in the Government of Canada, for assessing the impact of public programs.* You should recognize that there is considerable room for improvement for developing and using evaluation methodologies that are both well-suited to the reality of public programs and provide valid conclusions regarding impact.

The other issue emphasized in this paper was the necessity of producing evaluation reports that are relevant to the needs of decision-makers. An evaluation, in and of itself, has no inherent value; it is only valuable if it helps decision-makers make well-informed choices. Even when evaluations follow the highest standards of technical quality, their use to decision-makers will be undermined – perhaps negated – by evaluation reports that do not tell decision-makers, clearly, simply and succinctly, what they need to know.

Sources and Further Reading

Baker, Judy (2000). *Evaluating the Impact of Development Projects on Poverty. A Handbook for Practitioners.* Washington: The World Bank.

Bamberger, Michael (2006). *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints,* Washington: The World Bank.

Cummings, Rick (2006). " 'What if': the counterfactual in program evaluation," *Evaluation Journal of Australasia,* Vol. 6 (new series), No. 2, pp. 6-15.

Estrella, Marisol and John Gaventa (1998). *Who Counts Reality? Participatory Monitoring and Evaluation: A Literature Review.* Sussex: Institute of Development Studies.

Leeuw, Frans and Jos Vaessen (2009). *Impact Evaluations and Development. NONIE Guidance on Impact Evaluation.* Washington: NONIE.

OECD DAC Network on Development Cooperation (2010), *Evaluating Development Cooperation. Summary of Key Norms and Standards,* Paris: OECD

Pierce, Juliet (2004). "The search for the end of the rainbow – is impact evaluation possible?" IOD PARC, www.iodparc.com/resource/impact_assessment_possible.html

Purdon, Susan, et. al. (2001). "Research Methods for Policy Evaluation," Department of Work and Pensions Research Working Paper No. 2, London: Department of Work and Pensions

Ravallion, Martin (2001). "The Mystery of the Vanishing Benefits. An Introduction to Impact Evaluation," *World Bank Economic Review,* Vol. 15, No. 1, pp. 115-140.

Shadish, William R., Cook, Thomas D. and Campbell, Donald T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Treasury Board Secretariat of Canada, *Directive on the Evaluation Function,* www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15681&section=text

Treasury Board Secretariat of Canada, *Policy on Evaluation,* www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15024&section=text

Treasury Board Secretariat of Canada, *Standard on Evaluation for the Government of Canada,* www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15688&section=text

U.S. Department of Health and Human Services. Centers for Disease Control and Prevention. Office of the Director, Office of Strategy and Innovation (2005). "Introduction to program evaluation for public health programs: A self-study guide." Atlanta: Centers for Disease Control and Prevention.

Web Center for Social Research Methods, http://socialresearchmethods.net

White, Howard (2006). *Impact Evaluation – The Experience of the Independent Evaluation Group of the World Bank.* Washington: The World Bank.

White, Howard and Michael Bamberger (2008). "Introduction: Impact Evaluation in Official Development Agencies," *IDS Bulletin,* Vol. 39, No. 1, pp. 1-11.