Editor’s Remarks / Un mot du rédacteur  
Robert Schwartz  

ARTICLES  

Theory of Change Analysis: Building Robust Theories of Change  
John Mayne  

Théories du changement : comment élaborer des modèles utiles  
John Mayne  

Advancing Patient Engagement in Health Service Improvement:  
What Can the Evaluation Community Offer?  
Nathalie Gilbert and J. Bradley Cousins  

L’Ombudsman face aux défis de l’évaluation : est-il possible d’évaluer l’intangible?  
Seyive Wilfried Affodegon, Steve Jacob et Éric Montigny  

RESEARCH AND PRACTICE NOTES /  
NOTES SUR LA RECHERCHE ET LES MÉTHODES  

Making Evaluation More Responsive to Policy Needs: The Case of the Labour Market Development Agreements  
Yves Gingras, Tony Haddad, Andy Handouyahia, Georges Awad, and Stéphanie Roberge
Using Rubrics for an Evaluation: A National Research Council Pilot
Ghislaine Tremblay, Frédéric Bertrand, and Melissa Fraser

Moving Beyond the Buzzword: A Framework for Teaching Culturally Responsive Approaches to Evaluation
Ayesha S. Boyce and Jill Anne Chouinard

The Impact of Practice on Pedagogy: Reflections of Novice Evaluation Teachers
Jill Anne Chouinard and Ayesha S. Boyce

BOOK REVIEWS / COMPTES RENDUS DE LIVRES

Monitoring and evaluation training: A systematic approach, by S. G. Chaplowe and J. B. Cousins
Reviewed by / compte-rendu par Charles Lusthaus and Hussein Amery

Credible and actionable evidence: The foundations for rigorous and influential evaluations, by S. I. Donaldson, C. A. Christie, and M. M. Mark
Reviewed by / compte-rendu par Robyn Thomas Pitts, S. Austin Cavanaugh, and Emma Sunnassee

Instructions to Authors
Instructions aux auteurs
## Editorial Team / Équipe éditoriale

<table>
<thead>
<tr>
<th>Position</th>
<th>Name</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor-in-chief / Rédacteur en chef</td>
<td>Robert Schwartz, Ph.D.</td>
<td>Professor, Dalla Lana School of Public Health Executive Director, Ontario Tobacco Research Unit University of Toronto 155 College Street, Room 540 Toronto, ON, M5T 3M7 <a href="mailto:robert.schwartz@utoronto.ca">robert.schwartz@utoronto.ca</a></td>
</tr>
<tr>
<td>Associate Editor / Rédactrice associée</td>
<td>Isabelle Bourgeois, Ph.D.</td>
<td>Assistant professor, École nationale d'administration publique (819) 771–6095 ext. 2231 <a href="mailto:isabelle.bourgeois@enap.ca">isabelle.bourgeois@enap.ca</a></td>
</tr>
<tr>
<td>French Language Editor / Rédactrice associée francophone</td>
<td>Astrid Brousselle</td>
<td>Professeure titulaire, Département des Sciences de la santé communautaire Université de Sherbrooke (450) 466-5000, ext. 3682 <a href="mailto:astrid.brousselle@usherbrooke.ca">astrid.brousselle@usherbrooke.ca</a></td>
</tr>
<tr>
<td>Book Review Editor / Rédactrice, Comptes rendus de livres</td>
<td>Jill Anne Chouinard, Ph.D.</td>
<td>Assistant Professor, University of North Carolina at Greensboro <a href="mailto:jill.chouinard@uncg.edu">jill.chouinard@uncg.edu</a></td>
</tr>
<tr>
<td>Editorial Assistant / Adjointe à la rédaction</td>
<td>Emily Taylor</td>
<td>Canadian Journal of Program Evaluation University of Toronto (416) 978-8843 <a href="mailto:emily.taylor@camh.ca">emily.taylor@camh.ca</a></td>
</tr>
</tbody>
</table>
EDITORIAL BOARD / COMITÉ DE RÉDACTION

Courtney Amo A/Director, Evaluation and Risk Directorate; Atlantic Canada Opportunities Agency (ACOA), Government of Canada

Tim Aubry School of Psychology, University of Ottawa

Ilse N. Blidner Consultant, Toronto

Paul Brandon College of Education, University of Hawaii

Brad Cousins Faculty of Education, University of Ottawa

Nathalie Dubois École nationale d'administration publique

Paul Favaro Chief of Assessment and Accountability, Peel Board of Education

Swee Goh School of Management, University of Ottawa

Christophe Kellerhals Secrétaire permanent de la Commission externe d'évaluation des politiques publiques du canton de Genève (CEPP)

Marlène Laeubli Consultant, Switzerland

Chris Lovato University of British Columbia

Charles Lusthaus Consultant, Montréal

John Mayne Consultant, Ottawa

James McDavid School of Public Administration, University of Victoria

Céline Mercier Centre de réadaptation Lisette Dupras, Lachine

Anita Myers Department of Health Studies, University of Waterloo

Michael Obrecht Canadian Institutes of Health Research

John Owen Centre for Program Evaluation, University of Melbourne

Burt Perrin Consultant, France

Cheryl Poth University of Alberta

Hallie Preskill Executive Director, FSG Social Impact Consultants, San Francisco

Lucie Richard Faculté des sciences infirmières, Université de Montréal

Valéry Ridde Faculté de médecine, Université Laval

Ray C. Rist The World Bank, Washington, D.C.

Mark Seasons School of Planning, University of Waterloo

Robert V. Segsworth Department of Political Science, Laurentian University

Souraya Sidani Faculty of Nursing, University of Toronto

Nick L. Smith School of Education, Syracuse University

Sanjeev Sridharan Health Policy Management & Evaluation, University of Toronto

Rosalie Torres Developmental Studies Center, Oakland

Sue Weinstein Consultant, Toronto
It gives me great pleasure to introduce my last issue as Editor-in-Chief of CJPE. I feel that with this issue, I am going out with a nice splash.

CJPE has been at the forefront of publishing scholarly writing on the Theory of Change, and John Mayne has been a prolific contributor to this discourse. It is an honour to end my tenure at CJPE with an issue that contains two Theory of Change articles authored by John. The first introduces criteria and approaches for assessing theories of change. It follows on John’s 2016 best article of the year, published last year in English and reproduced here in French, which provides a structured framework for developing useful theories of change.

The two remaining full-length articles contribute new knowledge to specific evaluands. Gilbert and Cousins present a participatory evaluation approach to patient engagement and suggest ways in which the evaluation community can better engage in advancing effective patient engagement. Affodegon and Jacob seek to advance evaluation approaches for something quite different—the work of ombudsmen. Their critical review of evaluation approaches will undoubtedly prove helpful for future evaluations of rights protection agencies.

The Practice Notes section starts with two pieces that highlight innovative approaches to evaluation at the federal government level in Canada. Gingras et al. demonstrate the utility of small targeted studies and rich administrative panel data for evaluations in Employment and Social Development Canada. Tremblay, Bertrand and Fraser offer insights into the value of evaluation rubrics in a science organization.

Following on a recent CJPE Special Issue on cultural evaluation, Boyce and Chouinard present a framework for helping evaluators translate theoretical understanding of cultural evaluation into practice. A second Practice Note by the same authors similarly highlights translation of theory into practice—this time for students enrolled in an evaluation theory and practicum course.

It has been a great privilege to serve as Editor-in-Chief of CJPE. I have enjoyed the opportunity to learn so much from reviewing a wide variety of submissions and from guiding a good proportion to publication. Isabelle Bourgeois is so very well placed to take the baton and to lead CJPE through its next new and exciting phase.

Signing off,

Rob
C'est avec grand plaisir que je vous présente mon dernier numéro à titre de rédac-teur en chef de la RCÉP. J'ai l'impression de finir en beauté!

La RCÉP a été à l'avant-garde de la publication d'écrits scientifiques sur la théorie du changement et John Mayne a contribué de façon soutenue à la conver-sation sur le sujet. C'est un honneur pour moi de terminer mon mandat à la RCÉP avec un numéro proposant deux articles de John sur la théorie du changement. Le premier discute de critères et d'approches pour l'évaluation de théories du change-ment. Il vient compléter le meilleur article de John en 2016, publié l'an passé en anglais et reproduit ici en français, lequel offre un cadre structuré de mise au point de théories utiles du changement.

Les deux autres articles complets proposent de nouvelles connaissances d'objets d'évaluation précis. Gilbert et Cousins présentent une approche d'évaluation collaborative pour la participation des patients et suggèrent des moyens par lesquels la communauté de l'évaluation peut mieux faire la promo-tion d'une participation efficace des patients. Affodegon et Jacob cherchent à faire progresser les approches d'évaluation pour quelque chose de bien différent : le travail de protecteur du citoyen. Leur examen critique des approches d'évaluation en la matière sera sans doute d'une grande utilité pour les évaluations futures d'organisations de protection des droits.

La section des notes de pratique présente tout d'abord deux écrits qui mettent en relief des approches innovatrices en matière d'évaluation au gouvernement fédéral au Canada. Gingras et coll. démontrent l'utilité de petites études ciblées et de données administratives exhaustives recueillies au moyen d'un panel pour des évaluations chez Emploi et Développement social Canada. Tremblay, Bertrand et Fraser offrent une discussion intéressante sur la valeur de rubriques d'évaluation au sein d'une organisation scientifique.

Faisant suite à un récent numéro spécial de la RCÉP sur l'évaluation cul-turelle, Boyce et Chouinard présentent un cadre conçu pour aider les évaluateurs à traduire en pratique une compréhension théorique de l'évaluation culturelle. Une deuxième note de pratique par les mêmes auteurs met aussi en lumière la traduction de la théorie en pratique, cette fois pour les étudiants inscrits à un cours d'évaluation combinant théorie et stage.

Ce fut un très grand honneur pour moi que d'agir à titre de rédacteur en chef de la RCÉP. J'ai tiré beaucoup de plaisir d'avoir eu l'occasion d'améliorer mes connaissances au fil de mes lectures de diverses soumissions et du processus de
publication de bon nombre d’entre elles. Isabelle Bourgeois a tout ce qu’il faut pour prendre la relève et mener la RCÉP à la prochaine et fascinante étape de son développement.

Pour la dernière fois, je vous offre mes salutations.

Rob
Theory of Change Analysis:
Building Robust Theories of Change

John Mayne
Ottawa, Ontario

Abstract: Models for theories of change vary widely as do how they are used. What constitutes a good or robust theory of change has not been discussed much. This article sets out and discusses criteria for robust theories of change. As well, it discusses how these criteria can be used to undertake a vigorous assessment of a theory of change. A solid analysis of a theory of change can be extremely useful, both for designing or assessing the designs of an intervention as well as for the design of monitoring regimes and evaluations. The article concludes with a discussion about carrying out a theory of change analysis and an example.

Keywords: analysis of theories of change, criteria for good theories of change, impact pathways, theory of change

INTRODUCTION

Theories of change (ToCs) are now widely used in evaluations. They are the basis of theory-based evaluations (Coryn, Noakes, Westine, & Schroter 2011; Donaldson, 2007; Funnell & Rogers 2011; Rogers, 2007). As many have noted, the specific models used vary greatly (James, 2011; Valters, 2014; Vogel, 2012) and there is no overall agreement on just what comprises a ToC. Funnell and Rogers (2011, pp. 15–34) discuss the range of terms used and their histories, as does Patton (2008, pp. 336–340). Further, what constitutes a good or solid ToC is not at all clear; the characteristics or criteria of a robust ToC have not been widely discussed.

Corresponding author: John Mayne, john.mayne@rogers.com

© 2017 Canadian Journal of Program Evaluation / La Revue canadienne d'évaluation de programme
32.2 (Fall / automne), 155–173 doi: 10.3138/cjpe.31122
This article discusses criteria for a robust ToC and a tool for carrying out analysis of ToCs, namely Theory of Change Analysis (ToCA) to assess and strengthen ToCs.

When discussing specific aspects of ToCs and presenting examples, I will be using the behaviour change-based ToC model shown in Figure 1. Behaviour change-based ToCs are discussed in Mayne (2015) and the COM-B model in Mayne (2016a). However, the steps and principles discussed apply to theories of change generally.

**SOME TERMS**

Given the diversity of how terms around ToCs and results are used, let me first a review the terms being used here:

- *Results* is used to include outputs, outcomes, and impacts, where impacts are the final outcomes affecting well-being. A *result statement* is the exact text used to describe the result. The term *intervention* is used here to describe specific activities undertaken to make a positive difference in outcomes and impacts of interest. It covers policies, programs, and projects.

---

*The COM-B model postulates that behaviour (B) occurs as the result of interaction between three necessary conditions, capabilities (C), opportunities (O), and motivation (M).*
• **Impact pathways** describe causal pathways showing the linkages between a sequence of steps in getting from activities to impact. An intervention may have several pathways to impact.

• A **theory of change** (ToC) adds to an impact pathway by describing the causal assumptions behind the links in the pathway—what has to happen for the causal linkages to be realized. Theories of change are models of how change is expected to happen (ex ante case) or how change has happened (ex post case).

• **Rationale assumptions** identify the underlying hypotheses or premise(s) on which the intervention is founded.

• **Causal link assumptions** are the salient events or conditions necessary (or likely necessary) for a particular causal link in a ToC to be realized; if the assumption doesn’t hold, then the expected effect from that link will not occur. This can be a very demanding requirement, if interpreted literally. We can rather think in probabilistic terms, whereby causal link assumptions can be thought of as likely necessary assumptions, events, and conditions that almost always have to occur for the causal link to work.

Further discussion of these terms and other alternative terms used such as logic models and program theory can be found in Mayne (2015).

Because they are necessary or likely necessary, causal link assumptions also represent risks to the causal link occurring—the risk being that the assumption does not occur, that is, is not realized. For example, if an assumption is that local government takes some action, the risk is that it does not take the action. Consequently, rather than listing assumptions and risks, one can just identify assumptions.

Typically, some assumptions are less likely to be realized than others. For example, if an assumption is that some party, perhaps a local government, will take some action that has not been taken before, and nothing is being done to encourage the government to do so, then that assumption is quite likely at risk—and indeed may not be plausible. If an assumption is that a market will emerge for a new product and nothing is being done to encourage such a market, then that assumption is at risk. In addition, an assumption may be at risk because of counter pressures trying to ensure the assumption is not realized. An assumption that monitoring will be done by a third party may be at risk if there are other powerful parties who do not want the monitoring to be effective.

For theory of change analysis, I will call these at-risk assumptions. In an ex ante situation, at-risk assumptions represent potential gaps in the design of the intervention and likely serious threats to the intervention working. As a result, one may want to identify possible confirming actions that could be taken early on to give assurances that the assumption is likely to be realized, or corrective actions that might be taken to mitigate the at-risk assumptions. In an ex post situation, these are areas that need special attention in evaluations to see if in fact anything was done to address the risk.

In ex ante situations, it is important to keep the timeline in mind. At-risk assumptions for causal links well in the future may be less of a problem—realizing...
the risk, actions could be taken later to address the issue. Many assumptions would
not be expected to be at risk, such as when the assumption can be expected to be
realized based on past experience and/or research, if the intervention design is
solid, or even if it is agreed by stakeholders that it is likely to occur.

In a theory of change model, at-risk assumptions could be identified by bold-
ing the assumption and discussing it in the accompanying text.

**ISSUES IN ANALYZING A THEORY OF CHANGE**

In addition to their use in evaluations, ToCs have also been found useful in
designing interventions or assessing the designs of interventions (Leeuw, 2012;
Rey, Brousselle, & Dedobbeleer, 2012; Tremblay, Brousselle, Richard, & Beaudet
2013). Mayne (2015) and Mayne and Johnson (2015) discuss a variety of uses of
discuss some forms of analysis of ToCs. However, no structured approach for such
analysis has yet been proposed.

Those developing theories of change use forms of analysis both during de-
velopment and after. However, given the numerous elements of a ToC and the
various possible purposes, it is useful to undertake a structured analysis with
specific aims in sight. The *theory of change analysis* (ToCA) discussed here aims
at addressing two questions:

1. Does the intervention ToC appear robust? That is, is the ToC structurally
   sound and plausible?
2. What are the implications for monitoring and evaluating the intervention?

ToCA is done on a proposed ToC, one that has been developed to reflect how
an intervention is working or was expected to work; hence the “appear” term in the
question. Reality might suggest that the intervention and its ToC were not in fact
that robust! But *a priori*, before undertaking extensive data collection, we would
want to identify any evident shortcomings in the ToC and hence the intervention
design. And *ex post*, if we find that a ToC that has been used to model an interven-
tion is not very robust, we might find that helpful in explaining a less than successful
intervention and/or identifying issues that an evaluation should explore.

**Criteria for Robust Theories of Change**

When a ToC is being developed, the expectation is that it is not just a bunch
of ideas, but that it is well articulated, credible, plausible, and logical—that it
is robust. A robust ToC is defensible, would support a well-designed plausible
intervention design, and would provide a solid basis both for monitoring and for
theory-based evaluations.

A related idea is that of a ToC being evaluable, for which Rick Davies (2012) has
set out a list of criteria. Davies's criteria are quite broad in their coverage, meant to
include anything that is called a ToC. And indeed, as noted above, a wide range of models and representations have been used to depict theories of change, and several of Davies’s criteria challenge what has been set out as to whether it is a ToC at all.

My starting point is a little different. In defining above what a ToC is, it is assumed that what is being examined sets out the pathways of change as a causal sequence of results, and assumptions behind the pathways.

Thus, several of Davies’s criteria are “assumed,” namely testable, explained, complete, and inclusive. How well those criteria are addressed in a ToC is part of the robust criteria discussed below. Most of Davies’ other criteria are covered in the robust criteria as well. In addition, several other criteria are needed to assess the robustness of a ToC.

A robust ToC is one that is structurally sound and plausible. A robust ToC supports a solid and plausible intervention design: with this design, it is reasonable to expect that the intervention, if implemented as designed, will be able to contribute to the intended results. Criteria for a robust theory of change for an intervention would address the following questions:

For a structurally sound ToC:

1. Is the ToC understandable? Are there pathways of results, and are causal link assumptions set out? Is there a reasonable number of results?
2. Are the ToC results and assumptions well defined?
3. Is the timing sequence of results and assumptions plausible?
4. Is the ToC logically coherent? Do the results follow a logical sequence? Are the causal link assumptions pre-events and conditions for the subsequent effect? Is the sequence plausible or at least possible?
5. Are the causal link assumptions necessary or likely necessary?
6. Are the assumptions independent of each other (recognizing that some assumptions may apply for more than one causal link)?

For a structurally sound ToC that is plausible:

7. Is the ToC generally agreed?
8. Are the results and assumptions, or at least the key results and assumptions, measurable? What is the likely strength or status of evidence?
9. Are the causal link assumptions likely to be realized? Are at-risk assumptions mitigated through confirming or corrective actions?
10. Are the sets of assumptions for each causal link along with the prior causal factor plausibly sufficient to bring about the effect?
11. Is the level of effort (activities and outputs) commensurate with the expected results?
12. To what extent are the assumptions sustainable?
A ToC that is reasonably robust would provide a solid basis for using the ToC (Mayne & Johnson, 2015) in (a) designing and planning an intervention, (b) managing an intervention, (c) assessing and evaluating an intervention, and (d) scaling up an intervention. Robustness, as imagined here, is not a 0-1 variable. Meeting all the criteria could be quite demanding. Rather, in most cases, one would be improving a ToC over time, moving toward a more robust version.

There is evident need for an intervention to be plausible. At the outset, clear gaps or flaws in the design will most probably lead to a less successful intervention. Evaluability assessments are now seen as exploring the plausibility of intervention design with a view to improving the design and/or to identifying if it makes sense to undertake an evaluation (Davies, 2013; Peersman, Guijt, & Pasanen, 2015; Trevisan & Walser, 2014). The criteria here for a robust ToC include those used in evaluability assessments. The criteria also include those set out for SMARTly describing outcomes (Smart, Measurable, Achieved, Relevant, Timely) in Outcome Harvesting (Wilson-Grau & Britt, 2013).

**M&E Implications**

One purpose of a ToC is to provide a framework for setting out monitoring and evaluation plans. In carrying out ToC Analysis, it becomes clear just what needs to be monitored and paid attention to in evaluations. Questions here would be:

1. What data on results and assumptions should be monitored?
2. What issues need attention in an evaluation?
3. What is the likely strength or current status of evidence for the various results and assumptions, and in particular for each causal link?

Table 1 pulls these criteria together for ToCA. Each criterion is then discussed. These criteria build on ones I suggested earlier (Mayne, 2011).

**Ex Ante and Ex Post Perspectives**

In carrying out ToCA, it is important to keep in mind the perspective being used, namely if the situation is *ex ante* or *ex post*. The analysis is similar in both cases, but the implication of the findings will differ.

The context for the *ex ante* perspective is where a ToC is being developed for an intervention that has yet to be implemented or is in the early stages of implementation. The intent would be to develop a robust ToC to match a plausible intervention design, so that at the outset it seems reasonable that the intervention would bring about the expected results. In this setting, ToCA can be used to

- facilitate agreement on a ToC
- identify possible gaps in the intervention design and what can be done
• identify results and assumptions that need monitoring
• identify issues that a future evaluation needs to address.

In an *ex post* setting, the intervention has been in operation for some time and an evaluation is to be undertaken to see the extent to which the intervention has actually worked. Some monitoring data may have been gathered and some changes in the intervention may have been made over time. There is a need to either build (reconstruct) a ToC or revise an earlier ToC to reflect how the intervention is now seen as working. *Ex post* ToCA can be used to

• facilitate agreement on a robust ToC, often a reconstructed ToC
• identify current intervention design weaknesses that may explain limited expected results being achieved
• identify results and assumptions data that an evaluation needs to collect or get from monitoring data
• identify evaluation questions that need addressing in an evaluation.

Note that in this *ex post* scenario, ToCA itself would not be assessing if the ToC was in fact realized. That would be done as part of the evaluation, using something like contribution analysis.

In either case then, ToC Analysis would seek to

• *strengthen the ToC*: identify and correct any structural weaknesses in proposed theories of change.
• *strengthen the intervention design*: identify weakness in intervention design and what could be, or should have been, done to strengthen the design
• *identify data needs*: identify monitoring and evaluation data that need to be collected for assessing performance of the intervention.

**THE THEORY OF CHANGE ANALYSIS CRITERIA**

The criteria in Table 1 can be used to assess the robustness of the ToC and the underlying intervention. However, as noted earlier, robustness is not a 0–1 rating. That is, because there can be different models for the ToC of an intervention with different levels of detail, the criteria need to be applied in a sensible manner. They might best be thought of as guidelines for assessing the strength of a ToC and the intervention it represents.

**Overall Criteria**

*Understandable:* The ToC and especially its pathways should be clearly evident so that readers understand the intervention in the same way. I have argued elsewhere that a complex ToC needs to be unpacked into several nested ToC models (Mayne, 2015). Further, in any one ToC, there should be a reasonable number of results statements, so that the ToC model is “readable” to others beyond those
who developed it. A rule of thumb that I have used is that if you have more than 13 result “boxes,” you may have a mess instead of a ToC.

**Agreed:** If the ToC, no matter how well constructed, is the product of just one person or a small group, it may not have much support or buy-in, and may not be

---

**Table 1. Criteria for Theory of Change Analysis**

<table>
<thead>
<tr>
<th>Overall Criteria</th>
<th><strong>Understandable</strong></th>
<th>Is the logic and structure of the ToC clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Agreed</strong></td>
<td>To what extent is the ToC agreed or contestable?</td>
</tr>
<tr>
<td></td>
<td><strong>Level of effort</strong></td>
<td>Are the activities and outputs of the intervention commensurate with the expected results?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criteria for Each Result</th>
<th><strong>Well-defined</strong></th>
<th>Is the results statement unambiguous?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Plausible timing</strong></td>
<td>Is the time frame for the result reasonable?</td>
</tr>
<tr>
<td></td>
<td><strong>Logical coherence</strong></td>
<td>Does the result follow logically from the previous result? Is the sequence plausible or at least possible?</td>
</tr>
<tr>
<td></td>
<td><strong>Measurable</strong></td>
<td>Is there a need to measure the result? How can the results be measured? What is the likely strength or status of evidence for the result being realized?</td>
</tr>
<tr>
<td></td>
<td><strong>M&amp;E implications</strong></td>
<td>What are the implications for monitoring and evaluation?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criteria for Each Assumption</th>
<th><strong>Well-defined</strong></th>
<th>Is the assumption unambiguous?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Logical coherence</strong></td>
<td>Is the assumption a precondition or event for the effect sought?</td>
</tr>
<tr>
<td></td>
<td><strong>Justified</strong></td>
<td>What is the justification for the assumption as being necessary or likely necessary?</td>
</tr>
<tr>
<td></td>
<td><strong>Realized</strong></td>
<td>Is it plausible that the assumption will be realized? Are there at-risk assumptions that should be addressed?</td>
</tr>
<tr>
<td></td>
<td><strong>Sustainable</strong></td>
<td>Is the assumption sustainable?</td>
</tr>
<tr>
<td></td>
<td><strong>Measurable</strong></td>
<td>Is there a need to measure the assumption? How can the assumption be measured? What is the likely strength or status of evidence for the assumption being realized?</td>
</tr>
<tr>
<td></td>
<td><strong>M&amp;E implications</strong></td>
<td>What are the implications for monitoring and evaluation?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criteria for Each Causal Link</th>
<th><strong>Independence</strong></th>
<th>Are the assumptions for the link independent from each other?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>A sufficient set</strong></td>
<td>Are the set of causal link assumptions along with the prior causal factor sufficient to bring about the effect? Is the link plausible?</td>
</tr>
<tr>
<td></td>
<td><strong>Strength/Status of evidence</strong></td>
<td>What is the strength or current status of evidence for the causal link being realized?</td>
</tr>
</tbody>
</table>
sustainable in the sense of not being used or not lasting long. Broad agreement among stakeholders would usually suggest a more robust ToC, often built up through a participatory approach to building the ToC. And there may be different views on how the intervention is supposed to work. In this case, one may need to build more than one robust ToC and check each against reality in due course, See, for example, Hansen and Vedung (2010).

Level of effort: This is a rough check on the plausibility of the intervention. Does it seem reasonable that the activities of the intervention and their outputs will be enough for the intervention to make a difference in the ways expected? Interventions sometimes have quite ambitious intentions that are expected to be realized from a quite modest level of effort.

Criteria for the Results

Well defined: The results need to be as well defined as practical as to their meaning and content, and their measurability. They should not be subject to different interpretations by different readers.

Plausible timing: There should be an indication of when the results are expected to come about, and the time frame set out should be realistic, that is, plausible. Setting out realistic timing for when results can be expected is frequently neglected in developing ToCs, indeed often completely absent. Unrealistic expectations about timing can point to quite unrealistic interventions. Even less attention is paid to the trajectory of the expected results, as Woolcock (2009) discusses.

Logical coherence: That is, the step-by-step model from activities/outputs to impact should make sense, based on plausible or at least possible logic and perhaps prior evidence. The distinction here between plausible and possible logic reflects the fact that different ToC models provide different levels of detail. A “possible” logic sequence implies that the causal step is possible but represents a large leap in logic, which may be due to the level of detail in the ToC or to a causal link at-risk. Remember, the ToC is a model of expectations, which may of course turn out otherwise. If the ToC is behaviour-based, such as discussed by Mayne (2015) or Morton (2015) and illustrated in Figure 1, this significantly strengthens the logical coherence of the model.

Measurable: The results, or at least key results, should be measurable—there are reliable and valid measures of the results, and the needed data can be (readily) collected. Depending on the use being made of the ToC, there may not be a need to measure all the results set out in a pathway. For example, in a behaviour-based theory of change model, measuring capacity change can sometimes be a challenge. On the other hand, measuring behaviour changes is usually much simpler, and may be all that is needed if the expected behaviour changes have occurred and other aspects of the model are verified. The ToC analysis should indicate if the result (a) needs to be measured, (b) might be useful to measure, or (c) do not really need to be measured. It is useful here to note the likely strength of evidence based on the measures.

Implications for monitoring and evaluation: As part of the analysis, one can also assess what the implications of each component of the ToC are for monitoring and evaluation. Implications could be identifying evaluation questions to be addressed;
issues that need to be carefully watched or explored; issues, results, and/or assumptions that should be monitored; and/or identifying data that should be collected.

The analysis would identify specific M&E actions that should be taken to strengthen monitoring and evaluation.

Criteria for the Assumptions

Well defined: The events and conditions set out in the assumptions need to be as well defined as practicable as to their meaning, content, and measurability. They should not be subject to different interpretations by different readers.

Logical coherence: Because the assumption should be needed for the effect to occur, it should be a logical precondition or event.

Justified: The assumptions are justified by a solid argument as necessary or likely necessary events or conditions for the causal link to work.

Realized: One should expect that the assumptions would be realized. That is, there is general agreement, strong logic, actions being taken, or prior evidence that make the assumption plausible. The analysis could identify the at-risk assumptions that exist to the intervention and corrective or confirming actions that could or need to be taken to mitigate the risk. Ex ante, this would identify weaknesses in the intervention design and what might be done to strengthen the design. Ex post, it would identify assumptions that need careful examination in an evaluation. In essence, here one is assessing the degree of control the intervention has over the assumption.

Sustainable: An assumption may be realized during the period of the intervention, but one normally would hope that the assumption is sustainable after the intervention is over. If not, then the assumption is at future risk, as would be the causal link, the result in question, and indeed the intervention. Where sustainability is an issue, the intervention might want to undertake some form of corrective action.

Measurable: The assumptions, or at least key ones, are measurable: there are reliable and valid indicators, the relevant data can be (readily) collected, and/or there is adequate prior evidence. The analysis should indicate if the assumption (a) needs to be measured, (b) might be useful to measure, or (c) does not really need to be measured. Again, it is useful here to note the likely strength of evidence based on the measures.

The M&E Implications criteria are discussed above.

Criteria for Each Causal Link

Independence: For each causal link, the assumptions should be independent of each other—that is, be separate events/conditions—and hence be a minimum set of assumptions, recognizing that the same assumption may be needed for more than one causal link.

A sufficient set: The set of the initial result plus assumptions for the causal link should be seen as sufficient for that link to work, that is, for the cause plus assumptions to contribute to the effect. The link should be plausible—the link causal package should be enough to likely bring about the effect.

Strength/status of evidence: This final criterion is about the likely strength of the evidence on the causal link occurring (ex ante), or the current status of evidence.
about the link having been realized (ex post), classified as strong, medium, or weak. Again the analysis would seek to identify intervention design weaknesses or issues that need exploring in an evaluation. Where evidence appears weak, this might suggest the need for additional monitoring, research, and/or evaluation.

**CARRYING OUT TOC ANALYSIS**

The actual ToC Analysis needs to carried out in a step-by-step manner. Too often, a theory of change is developed on the basis of the ideas and beliefs of those involved without much challenge and analysis. Without structured analysis and challenge, it is unlikely that a robust theory of change and the implications for intervention design would emerge. ToCA entails a careful examination of each element in the theory of change, how the elements fit together, and an assessment of the ToC weaknesses, data needs, and their implications.

ToCA would use the criteria in Table 1 as the basis for analysis, roughly in the order set out. There would likely be some interplay among results, assumptions, and the pathway. The findings could then be summarized in terms of implications around the two questions noted earlier.

**Step 1: Overall Criteria**

The initial analysis is to determine if there is indeed an actual ToC model to work with, and if the intervention seems at all plausible.

*Understandable:* If the ToC is hard to understand, such as if pathways are unclear or there is a proliferation of results, then rethinking and redrafting are needed so that there is something resembling a ToC with impact pathways.

*Agreed:* If there are different views as to how the intervention is expected to work, more discussion on the ToC is probably warranted. If differences persist, then it may be necessary to build more than one ToC and analyze each of them.

*Level of effort:* If the expectations for results are quite out of line with the level and nature of the activities being undertaken, there may be a need to rethink the design of the intervention or to reduce expectations to a more realistic level.

**Step 2: Detailed ToC Analysis**

The detailed ToC Analysis is best done result level-by-result level in sequence. That is, using the behaviour-based ToC model, in order:

- **Getting to Reach:** Will the outputs delivered reach the intended target groups with the right reaction?
- **Getting from Reach to Capacity Change:** Will the outputs delivered and their reach lead to the intended capacity changes?
- **Getting from Capacity Change to Behaviour Change:** Will the capacity change lead to the intended Behaviour Changes?
- **Getting from Behaviour Change to Direct Benefits:** Will the behaviour changes lead to the intended Direct Benefits?
• Getting from Direct Benefits to Well-being Changes: Will the direct benefits lead to the intended Well-being Changes?

If another ToC model is being used, the steps are the same: getting from one level to the next. For each level, an analysis of results and an analysis of assumptions would be done and a summary of findings set out:

Analysis of Results

Definition: If not well defined, need to further define terms.
Timing: If not sensible, suggests a structural change to the ToC needed.
Logical coherence: If not OK, suggests a structural change to the ToC needed.
Measurement: Indicate if needed, might be needed, not needed, and through what means. If strength of evidence is weak and the measurement important, suggests an issue to be addressed in the M&E Implications.

M&E implications: Brings together the M&E issues. Need to remember that not all results may need to be measured.

Analysis of Assumptions

Definition: If not well defined, need to redefine terms.
Logical coherence: If not OK, suggests the need for structural changes in the ToC.
Justification: If not necessary or not likely necessary, then the assumption should be dropped.
Realization: If realization is in doubt, then need to identify assumption as at-risk and set out confirming or corrective actions.
Sustainability: Similarly, if the assumption is found not to be sustainable, a corrective action may be needed, or, in an ex post case, the issue noted as a lesson learned for future similar or follow-up interventions.
Measurable: Indicate how measures would be taken and if needed, might be needed, not needed. If strength of evidence is weak and measurement important, suggest an issue to be addressed in the M&E Implications.

M&E implications: Need to remember that not everything need be measured.

Assessing the Causal Link

Independence: If assumptions are not independent, consider merging assumptions.
A sufficient set: If not a sufficient set, additional assumption(s) or more of the prior result are needed.
Strength/status of evidence: Indicate level of evidence for the link being realized.

Summary of Findings for Getting to a Result

The summary of the analysis can depend on the specific purpose and context, but in general can highlight (a) the changes needed to enhance the
robustness—structural soundness and plausibility—of the ToC, (b) the level of evidence there is on results and assumptions, (c) the actions that are needed to enhance the robustness of the intervention design, and (d) the M&E implications.

Structural Changes Needed
Where the structural criteria for a robust theory of change are not met, structural changes are needed to the ToC to enhance its robustness—that is, changes in descriptions used, result statements, coherence, assumptions, and/or causal links. After any structural changes, we would want to conclude that the ToC is reasonably sound.

Strength/Status of Evidence
Summary analysis can indicate the strength of evidence for (a) the result in question, (b) the assumptions associated with the link, and (c) the link being realized.

Additional Intervention Effort Needed to Enhance Plausibility
If plausibility or sustainability of the ToC/intervention design is questionable due to at-risk assumptions identified and/or sustainability being questioned, then confirming or corrective actions are likely needed. We would want the analysis to conclude that, with the confirming/corrective actions, the intervention design is (or ex post would be) robust. Where the ToC is seriously contested, more than one ToC may needed to be developed and analyzed.

M&E Actions
To monitor how well implementation is going or to verify the ToC in an evaluation, it is important to identify what data need collecting and the likely strength of the resulting evidence.

Conclusion: Overall conclusions for the specific link (component) in the ToC on robustness, level of evidence, and sustainability.

Theories of change are best developed in a participatory manner involving those designing/implementing the intervention and the evaluator (Mayne, 2015, pp. 137–138). During this development, of course, the criteria for a robust theory of change can be kept in mind. In other cases, the ToC Analysis is done on a completed theory of change, probably (although not necessarily) by the evaluator. The findings of the analysis should then be discussed with intervention implementers. This discussion may bring to light issues that were not, but need to be, included in the ToC, identify issues about the intervention design that need addressing, and/or identify data that need to be monitored or that need to be addressed in a planned evaluation.

AN EXAMPLE
To illustrate issues and concepts in ToC Analysis, I examined a previously used case of an intervention aimed at improving the nutritional diets of children through providing training and information to mothers (Mayne, 2015). The ToC used there
is shown in Figure 2, with a few small changes to be consistent with the COM-B model: the motivation and food availability assumptions have been shown as capacity change assumptions rather than behaviour change assumptions.

The ToC in Figure 2 was used to carry out the ToCA. All the details are not provided here—but can be found in Mayne (2016b), as the analysis is quite lengthy. And that is worth a note. ToC Analysis is not a quick and dirty approach: it takes time, but not a lot, and patience to go through each criterion for each result and each assumption. But it can be worthwhile. Having developed the original example, I was not expecting many new insights, but I was wrong!

The findings from the ToCA are summarized below for each results level.

**Figure 2. A Nutrition Intervention Theory of Change**

**Getting to Reach and Reaction**

Several needed *structural changes* were identified. The Reach result statement was not well-defined. What did “mothers with young children reached” mean? It could mean several things, such as mothers heard about the training, mothers were asked to participate, or mothers participated in at least the first session. I
assumed it was the latter case: reach and reaction was asking if mothers at least started the training and if they had a positive initial reaction. So the result statement needed to be changed.

Further, the first reach assumption was the same as the reach statement! Clearly there was a logical coherence problem. A new assumption, in fact two, were needed: targeted mothers are well identified, and targeted mothers can be communicated with. To get participation, the intervention needed to know who and where the targeted mothers were, and needed to be able to get the message to them about the nutrition training sessions.

If sustainability in the target area was an issue, then there would need to be a plan of how new mothers beyond the initial reach were to be reached, such as perhaps building into the training the need to spread the word within their communities.

There are two M&E implications: namely, the need to track the percentage of targeted population that initially participated, and to monitor initial reaction of participants.

**Getting from Reach to Capacity Change**

Several small structural changes were needed in the wording of the capacity result and assumptions (see Figure 3, where the changes are underlined).

Assumption 2 about the availability and affordability of nutritious food (opportunities) is possibly at-risk without more information. In Figure 3, at-risk assumptions are bolded. A useful corrective action would be to make local markets aware of the intervention and the expected increased demand for certain food products.

And a confirming M&E action is needed: the availability and affordability of nutritious food should be monitored during the life of the project.

**Getting from Capacity Change to Behaviour Change**

Behaviour change assumptions 1 and 2 overlap somewhat and may be at-risk. The intervention may need a better understanding about how decisions on food are made in households, and the sessions offered to households rather than only mothers.

*Me&E implications:* Household surveys could track adoption of the new practices and general household support, and identify problems. Perhaps schedule a survey after 2 months and a follow-up 1 year later.

**Getting from Behaviour Change to Direct Benefits**

Assumption 2 about substituting other foods is at-risk. A confirming action could be to include this substitution issue in the nutrition training.

*Me&E implications:* Follow-up household surveys could track children's dietary intake.

Overall, although implied by the timeline, Figure 2 did not set out a clear time frame for the intervention to have an impact. The level of evidence on realizing the ToC would be good, using the measures suggested.
Based on this ToC Analysis, the revised and more robust ToC is shown in Figure 3. At-risk assumptions are shown in bold, and wording changes are underlined.

**CONCLUDING REMARKS**

Theories of change are the basis for theory-based evaluation approaches, such as logical analysis (Brousselle & Champagne, 2011; Rey et al., 2012), realist evaluation (Blamey & Mackenzie, 2007; Pawson, 2013), contribution analysis (Mayne, 2012), and process tracing (Schmitt & Beach, 2015). As such, the robustness of the theory of change used matters. A weak theory of change can only generate weak findings. For example, confirming a weak theory of change—one poorly structured with evident logical gaps—in contribution analysis cannot lead to credible causal contribution claims.

This article argues the usefulness of building robust theories of change and structured theory of change analysis, so that evaluation findings based on these theories of change are strengthened. ToC Analysis involves assessing a theory of
change against a set of criteria (Table 1) for each result, each assumption, and each causal link, challenging the structure and logic of the theory of change. The analysis takes some time and discipline to carry out. But it is mainly a desk review, and overall it entails hours rather than days of work. In my experience, it inevitably leads to improvements in the theory of change. The results are usually quite informative, leading to

- more robust ToCs,
- better intervention designs,
- useful M&E actions to help manage the intervention and support future evaluation, and
- \textit{ex post}, more credible theory-based evaluations.

REFERENCES


Tremblay, M.-C., Brousselle, A., Richard, L., & Beaudet, N. (2013). Defining, illustrating and reflecting on logic analysis with an example from a professional development program. *Evaluation and Program Planning, 40*, 64–73. [http://dx.doi.org/10.1016/j.evalprogplan.2013.05.004](http://dx.doi.org/10.1016/j.evalprogplan.2013.05.004)


**AUTHOR INFORMATION**

John Mayne is an independent advisor on public sector performance. Over the past 13 years he has focused largely on international development evaluation and results-based management work. He has been working with several government, NGOs, and international organizations. He has authored numerous articles and reports, including several on contribution analysis, and co-edited eight books on program evaluation and performance monitoring. In 1989 and in 1995, he was awarded the Canadian Evaluation Society Award for Contribution to Evaluation in Canada. In 2006, he was made a Canadian Evaluation Society Fellow. Dr. Mayne’s current research interests are on approaches for strengthening impact evaluation, useful theories of change, and dealing with attribution.
Théories du changement :
comment élaborer des modèles utiles

John Mayne
Ottawa (Ontario)

Résumé : Bien que les théories du changement soient souvent l’objet d’articles en évaluation et qu’il y ait un consensus sur ce qu’est, conceptuellement, une théorie du changement, il n’y a, en réalité, au-delà des principes généraux, que peu de consensus sur ce qu’elles comprennent, sur ce qu’elles devraient mettre en valeur ainsi que de la manière dont elles devraient être représentées et utilisées. Cet article donne un aperçu de certains modèles de théories du changement qui se sont révélés utiles, tant pour des interventions simples que complexes, et présente la façon dont ils ont été développés. Les modèles sont intuitifs et flexibles, leurs composantes sont bien définies et ils s’appuient sur des modèles de causalité rigoureux. Ces modèles fournissent un cadre structuré pour l’élaboration de théories utiles du changement et pour l’analyse de l’intervention.

Mots clés : liens de causalité, intervention complexe, chaîne d’effets, modèle logique, chaîne des résultats, théorie du changement, théorie de la portée.

Abstract: Although theories of change are frequently discussed in the evaluation literature and there is general agreement on what a theory of change is conceptually, there is actually little agreement beyond the big picture of just what a theory of change comprises, what it shows, how it can be represented, and how it can be used. This article outlines models for theories of change and their development that have proven quite useful for both straightforward and more complex interventions. The models are intuitive, flexible, and well-defined in terms of their components, and they link directly to rigorous models of causality. The models provide a structured framework for developing useful theories of change and analyzing the intervention they represent.

Keywords: causal links, complex intervention, impact pathways, logic model, results chain, theory of change, theory of reach

Les modèles décrivant la logique des interventions sont fréquemment analysés et utilisés dans le domaine de l’évaluation. Voir, par exemple, Patton (2008); Chen (2015); Rossi, Lipsey et Freeman (2004); Morra Imas et Rist (2009); Funnell et

J'emploierai ici le terme théorie du changement. Les théories du changement peuvent être utilisées de façon très variée, tant pour le développement, la gestion et l'évaluation des interventions. Mayne et Johnson (2015) discutent de l'utilisation des théories de changement pour :

La conception et la planification des interventions
1. Conception des interventions
2. Compréhension et représentations communes des interventions par les parties prenantes
3. Identification des enjeux en lien avec l'équité, le genre et les inégalités de pouvoir (empowerment issues)
4. Évaluation ex ante des interventions proposées

La gestion des interventions
5. Conception de systèmes de monitorage
6. Compréhension de la mise en œuvre, gestion adaptative et apprentissage

L'évaluation des interventions
7. Conception des questions, méthodes et outils d'évaluation
8. Formulation des hypothèses pour la production des effets
9. Production de rapports de performance

La transérabilité
10. Généralisation théorique, à d'autres contextes et à des échelles différentes.


une prolifération d’interprétations différentes sur ce qu’une théorie du changement comprend en pratique, sur la façon d’en élaborer une et de la décrire.


**THÉORIES DU CHANGEMENT ET CHAÎNE D’EFFETS**

Commençons par définir quelques termes clés. Le terme *résultats* englobe les effets proximaux, les effets intermédiaires ainsi que les impacts, les impacts étant les derniers résultats avant le bien-être. J’emploie le terme *intervention* pour décrire des activités précises mises en œuvre afin de produire une différence souhaitable au niveau des effets. Le terme intervention est employé autant pour les politiques, les programmes et les projets.

Pour comprendre si une intervention fonctionne et comment elle fonctionne, nous devons comprendre comment les activités de l’intervention sont censées produire les résultats escomptés, c’est-à-dire : a) le cheminement causal entre les activités et les résultats depuis les effets les plus proximaux jusqu’aux impacts; b) les hypothèses causales qui indiquent pourquoi et dans quelles conditions les liens causaux sont supposés fonctionner. De nombreux termes sont utilisés pour désigner ces liens causaux : chaîne des résultats, modèle logique et chaînes d’effets. J’emploierai le terme « chaîne d’effets ».

Une chaîne d’effets désigne les liens de causalité entre les diverses étapes depuis les activités jusqu’aux impacts. Une théorie du changement ajoute à la chaîne d’effets les hypothèses causales qui sous-tendent chacun de ses liens, c’est-à-dire qu’elle explicite ce qui doit se produire pour que les liens de causalité se concrétisent. Patton (2008, p. 336) fait la même distinction entre les modèles logiques et les théories du changement : « Le fait de préciser les mécanismes de causalité transforme un modèle logique en une théorie du changement. » Chen (2015), au chapitre 3, fait une distinction similaire. Les théories du changement modélisent la façon dont le changement est censé se produire (cas *ex ante*) ou s’est produit (cas *ex post*).

Il y a bien des façons de dépeindre les chaînes d’effets et les théories du changement, comme l’illustrent d’ailleurs Funnell et Rogers (2011). La figure 1 représente une théorie du changement générique de base qui s’est avérée utile dans plusieurs contextes. La séquence des cases de la figure est la chaîne d’effets (chaîne de résultats), dont il sera d’abord question. La figure 1 est une version améliorée du modèle de théorie du changement présenté dans Mayne (2014).
Figure 1. Théorie du changement générique et de base

Les composantes d’une chaîne d’effets

Les bénéficiaires sont les groupes cibles pour lesquels l’intervention vise à améliorer le bien-être. Ces groupes sont peut-être segmentés selon le niveau de revenu, le sexe, l’origine ethnique ou la région géographique. Prenons par exemple une intervention qui vise à améliorer la nutrition des enfants; les bénéficiaires visés sont les enfants. À noter qu’un groupe cible peut comprendre des organisations.

Dans le modèle de la théorie du changement, les activités et les résultats (des résultats proximaux depuis les produits jusqu’aux impacts) sont représentés dans les cases :

1. Produits de l’intervention (extrants)
2. Activités
3. Portée et réaction
4. Changements au niveau des capacités : savoir, attitudes, aspirations, compétences, opportunités
5. Changements de comportement
6. Avantages directs
7. Changement au niveau du bien-être
8. Hypothèses quant aux changements de comportement
9. Hypothèses quant aux avantages directs
10. Hypothèses quant aux changements de comportement
11. Hypothèses quant aux avantages directs
12. Hypothèses quant à la portée

Ligne du temps

Influences extérieures

Résultats imprévus
• Les activités sont les actions entreprises par les personnes qui effectuent l’intervention.
• Les produits de l’intervention, biens et services, sont les entrants qui résultent des activités. Dans notre exemple de la nutrition, il pourrait s’agir du matériel éducatif sur les bienfaits d’une saine alimentation. Dans cet article, le terme produit désigne ces biens et services directs.
• Les changements dans les capacités sont les changements liés aux savoirs de ceux qui ont reçu ou utilisé les biens et services de l’intervention, à leurs attitudes, compétences, aspirations et aux opportunités. Comme nous le verrons plus loin, tous ces changements sont nécessaires pour qu’une nouvelle action3 soit entreprise.
• Les changements de comportement sont les changements qui se produisent dans les pratiques réelles, lorsque les membres du groupe cible à rejoindre agissent différemment ou utilisent les produits de l’intervention. Dans notre exemple, il pourrait s’agir des changements dans les pratiques alimentaires des mères qui résulteront de l’amélioration des savoirs découlant de la formation. Il y a généralement une rétroaction entre la capacité et les changements de comportement (comme dans l’acquisition de nouveaux savoirs et de nouvelles compétences par l’action).
• Les avantages directs sont les améliorations de l’état individuel des bénéficiaires. Il peut notamment s’agir d’une augmentation du revenu, d’une utilisation accrue des services de santé, de pratiques agricoles plus productives, de l’autonomisation des bénéficiaires ou, dans notre exemple, de la consommation d’aliments plus sains par les enfants.
• Les changements au niveau du bien-être4 se traduisent par une amélioration cumulative à long terme du bien-être général des bénéficiaires : l’amélioration de la santé, la réduction de la pauvreté, une sécurité alimentaire accrue. Dans notre exemple, l’amélioration du régime alimentaire contribuerait à bonifier la nutrition et l’état de santé des enfants.

À noter que le cheminement causal présenté à la figure 1 ne présente pas la séquence de résultats en effets proximaux, intermédiaires et finaux comme c’est
habituellement fait. Comme ces termes d’usage courant n’ont guère de signification intuitive, ils ne sont pas très éclairants pour établir la chaîne d’effets et leur utilisation ouvre souvent la porte à des débats stériles, à savoir si, par exemple, il s’agit d’un résultat proximal ou intermédiaire. Et même si je pensais que le terme « extrants » était bien défini et généralement accepté, le récent guide du Groupe des Nations Unies pour le développement (2011) ajoute de la confusion en définissant les extrants comme des produits ou des changements de capacité. Je soutiens que la figure 1 est une représentation plus utile de la chaîne d’effets que le modèle courant.

Les influences extérieures sont les événements et conditions sans lien avec l’intervention qui pourraient affecter l’atteinte des résultats escomptés. Il peut s’agir d’autres interventions aux visées similaires ou encore de grandes tendances économiques ou sociales. Elles ne font pas partie inhérente de la théorie du changement relative à l’intervention. Dans notre exemple sur la nutrition, une baisse du prix des légumes pourrait aussi contribuer en partie à une augmentation de la consommation de légumes sans lien avec l’intervention éducative. L’enrichissement d’aliments industriels comme le sucre ou la farine pourrait également contribuer à expliquer une amélioration de la consommation de micronutriments.

La figure 1 indique les effets imprévus, soit les effets non anticipés positifs ou (plus souvent) négatifs qui résultent des activités de l’intervention. Si ces effets sont probables, il faut les représenter. Il importe de rechercher activement les effets imprévus ex post. Notons aussi que malgré son apparence linéaire, la figure 1 prévoit explicitement une non-linéarité au moyen de la rétroaction entre les divers stades. La figure 1 illustre aussi une ligne du temps indiquant à quel moment on peut s’attendre à ce que les changements escomptés se produisent. Une ligne du temps, même avec des dates approximatives, est un ajout utile à la chaîne d’effets.

**De la chaîne d’effets à la théorie du changement**

La première étape de l’élaboration d’une théorie du changement consiste à développer la chaîne d’effets. Mais une chaîne d’effets ou un modèle logique n’est pas une théorie du changement. C’est en ajoutant les hypothèses aux liens de causalité indiqués dans la chaîne d’effets que nous obtenons une théorie du changement. Ces hypothèses illustrées par des cases pointillées dans la figure 1 indiquent les principaux événements et conditions qui doivent se produire pour que chaque lien de la chaîne des effets fonctionne comme prévu. Qu’est-ce qui est nécessaire au fonctionnement du lien de causalité? Quels facteurs sont essentiels à ces processus de causalité? Pour des raisons pratiques, il nous suffit de considérer les principales hypothèses, celles qui ressortent du lot pour une raison ou une autre, qui sont remarquables et qui sont pertinentes pour la situation. D’autres, comme le fait que le soleil se lève tous les matins ou qu’une révolution ne s’est pas produite, ne sont pas pertinentes, bien qu’ex post, une révolution expliquerait facilement pourquoi l’intervention n’a pas fonctionné! L’articulation d’hypothèses sur les liens de causalité implique un savant mélange d’évidences scientifiques, d’expérience des parties prenantes et de théories provenant des sciences sociales. Par exemple,
dans le cas de l'alimentation des enfants, une hypothèse serait que le mari et la belle-mère approuvent les choix de la mère quant à l'alimentation des enfants et qu'ils la laissent prendre les décisions à ce chapitre.

Ces hypothèses sur les liens de causalité couvrent tous les risques associés à chaque lien de causalité; chacune des hypothèses est un risque pour la concrétisation de la théorie du changement\(^5\). Dans l'exemple de la nutrition, les risques relatifs à la disponibilité d'aliments nutritifs et à la modicité de leur prix seraient pris en compte dans l'hypothèse que des aliments nutritifs sont disponibles à prix modique.

- **Hypothèses quant à la portée**: Ces hypothèses sont les événements et conditions qui doivent être réunis pour que les produits de l'intervention (extrants) rejoignent les groupes cibles et soient reçus positivement par eux. Par exemple, on posera l'hypothèse que les produits de l'intervention rejoindront effectivement l'auditoire cible et que celui-ci les considérera comme acceptables et dignes d'intérêt. Ici, un des principaux risques est que le groupe cible ne soit pas le « bon » groupe, comme dans une intervention en nutrition infantile qui vise des mères alors qu'en fait, ce ne sont pas elles qui décident qui mangera quoi; il s'agit aussi de rejoindre la totalité du groupe ciblé et non, par exemple, uniquement les personnes qui sont volontaires pour participer.

- **Hypothèses quant aux changements dans les capacités**: Ces hypothèses sont les événements qui doivent se produire et les conditions qui doivent changer pour que les produits (extrants) qui atteignent les populations cibles induisent des changements au niveau de leurs savoirs, leurs attitudes, leurs compétences, leurs aspirations et de leurs opportunités – bref, leur capacité de faire les choses autrement. Par exemple, il pourrait s'agir que les produits soient compris, réalistes, culturellement acceptables, considérés comme utiles, cohérents avec les compétences et valeurs préalables de la population cible, considérés comme pertinents par le groupe visé.

- **Hypothèses quant aux changements de comportement**: Ces hypothèses sont les événements et conditions à réunir pour que les changements de capacité des groupes cibles se traduisent par une évolution réelle de leurs pratiques. On pourrait y inclure, entre autres, la capacité financière de réaliser les changements de pratique, l'acceptation par autrui (les pairs, les dirigeants sociaux, culturels et religieux, la famille) nécessaire à la réalisation des changements, la démonstration de l’utilité des changements de pratique, des politiques ou un cadre naturel favorables à l’adoption des pratiques, et l'accès aux biens et fournitures nécessaires.

- **Hypothèses quant aux avantages directs**: Ces hypothèses sont les événements et conditions nécessaires pour que les changements de pratique procurent un avantage direct sur les conditions des bénéficiaires ciblés. On pourrait notamment poser l’hypothèse qu’un changement de pratique
entraîne une augmentation nette du revenu, une utilisation régulière des services de santé ou une participation à la prise de décision. Dans le cas de la nutrition, on peut poser l’hypothèse que le seul changement réalisé dans le régime alimentaire est celui que recommande le programme de formation. Par contre, si les pratiques améliorées (p. ex. consommer davantage de légumes) sont intégrées, mais que d’autres apports nutritifs sont réduits, l’avantage attendu pourrait ne pas se concrétiser.

- **Hypothèses quant aux changements sur le bien-être** : Ces hypothèses sont les événements et conditions préalables nécessaires pour que les avantages directs entraînent des changements au niveau bien-être des bénéficiaires. Par exemple, si les enfants consomment des aliments plus sains et ont accès à des soins de santé de base et à de meilleures conditions d’hygiène, leur nutrition et leur état de santé s’amélioreront. Si, à la suite de l’intervention, les femmes commencent à jouer un rôle accru dans les décisions de consommation et que l’intervention est considérée comme réussie, ce changement pourrait contribuer à une évolution des normes de genre propre à renforcer le pouvoir d’agir des femmes.

On notera qu’une hypothèse sur le lien de causalité n’est pas une description du lien de causalité. La description d’un lien de causalité (ces liens sont représentés par des flèches continues dans la figure 1) serait, par exemple, que l’évolution des savoirs, des compétences et ainsi de suite (la capacité) se traduiront, en pratique, par les changements attendus de comportement. Les hypothèses sur les liens de causalité expliquent comment et pourquoi le lien de causalité fonctionne.


Dans l’exemple de la nutrition, on peut supposer sans risquer de se tromper que les mères veulent améliorer la santé de leurs enfants (motivation) et que l’intervention vise à leur en donner la possibilité et la capacité.

Il peut être utile de distinguer deux types de changements de capacité et de comportement. Les premiers sont *incrémentaux*; ce sont des changements marginaux à l’état actuel, par exemple l’apprentissage de techniques et de compétences...
nouvelles ou l'adoption de nouvelles pratiques. Ces changements sont plus faciles à créer que des changements fondamentaux, tels que le fait de penser les problèmes différemment ou de modifier ses pratiques actuelles. Dans notre exemple, si l'intervention vise l'acquisition de nouveaux produits alimentaires pour les enfants, il s'agit d'une pratique ajoutée et d'un objectif relativement simple. En revanche, si le changement de pratique recherché est une redistribution des aliments entre les membres du ménage, on s'attaque au mode actuel de distribution des aliments, ce qui souleve des enjeux de pouvoir. Ce changement fondamental sera vraisemblablement beaucoup plus difficile à implanter, de sorte que les hypothèses sur les liens de causalité qui y sont associées doivent être particulièrement robustes et ambitieuses.

Jusqu'ici, notre discussion est plutôt déterministe (par exemple, une hypothèse est nécessaire ou elle ne l'est pas). Mais nous pourrions aussi chercher à refléter la nature probabiliste de la causalité. Mahoney (2008, p. 421) avance qu'« un traitement est une cause lorsque sa présence augmente la probabilité qu'un résultat se produise dans n'importe quel cas ». Il présente les notions utiles de causes nécessaires en termes probabilistes (probabilistically necessary causes), soit des « facteurs qui doivent généralement ou presque toujours être présents pour que le résultat se produise », et de causes suffisantes en termes probabilistes (probabilistically sufficient causes), au sens d'une « cause qui, la plupart du temps, produit l'effet d'elle-même » (p. 425–426). Pour bien des interventions visées par une évaluation, il s'agit d'interprétations réalistes des notions de nécessité et de suffisance.

Ainsi, on peut envisager les hypothèses sur les liens de causalité comme des hypothèses, des événements et des conditions probablement nécessaires qui, presque toujours, doivent se produire pour que le lien de causalité fonctionne.

Énoncer des hypothèses associées à une théorie du changement peut être difficile, en raison de la diversité des types d'hypothèses associés à une intervention. En particulier, outre les hypothèses sur les liens de causalité dont il a été question plus haut, il existe aussi des hypothèses de rationalité, qui décrivent l'hypothèse ou prémisse sous-jacente servant de fondement à l'intervention. Par exemple, on présume que le fait d'informer les adultes d'un ménage des bienfaits d'une bonne nutrition pour leurs enfants entraînera un changement de comportement de leur part et une amélioration du régime alimentaire des enfants. On présuppose que la justification d'une intervention repose généralement sur des données probantes et sur l'expérience antérieure.

La figure 2 illustre la théorie du changement associée à l'exemple de la nutrition. Cet exemple n'est pas fondé sur un cas réel, mais White (2009) analyse une intervention très semblable menée au Bangladesh.

J'avance que, dans la plupart des interventions, chacune des composantes de la théorie du changement générique – les activités, les produits, les changements dans les capacités, les changements de comportement, les avantages directs, les changements au niveau du bien-être, de même que les hypothèses sur les liens de causalité associées – peuvent et doivent être déterminées et faire l'objet d'une
réflexion lors de l’élaboration de chaînes des effets et de théories du changement. La structure du modèle nous force à considérer la façon dont on s’attend à ce que les résultats escomptés surviennent : Quel est le processus de causalité à l’œuvre? Que faut-il pour qu’il se concrétise? *Le modèle est un cadre pour l’analyse du fonctionnement d’une intervention.*

**THÉORIES DU Changement ET FAISCEAUX DE CAUSALITÉ**

Les théories du changement illustrent comment et pourquoi on s’attend à ce qu’une intervention contribue à l’atteinte des effets escomptés. Toutefois, pour qu’ils se produisent, il faut plus que la simple réalisation des activités; il faut aussi que les hypothèses sur la causalité se manifestent. Les activités sont rarement la seule cause d’un effet. La théorie du changement décrit un *faisceau de causalité* constitué d’activités et d’hypothèses qui, ensemble, sont censées contribuer à
l’obtention des résultats escomptés (condition de suffisance). Cartwright et Hardie (2012) appellent ces hypothèses des facteurs de soutien (support factors), c’est-à-dire des événements et conditions essentiels pour contribuer à l’obtention des effets. On s’attend aussi à ce que les activités constituent une portion essentielle de ce faisceau de causalité (causal package) – condition de nécessité. Autrement dit, sans les activités, la manifestation des hypothèses sur les liens de causalité ne suffirait pas à produire l’effet escompté. On dit alors que les activités sont une cause contributive des résultats. Selon ces termes, une théorie du changement est un modèle de contribution causale de l’intervention; c’est un modèle qui indique l’ensemble des liens de causalité et leurs contributions à l’obtention des résultats. Mayne (2012) aborde la question des causes contributives et les faisceaux de causalité dans le contexte des théories de la causalité et plus particulièrement des causes INNNS7.

La théorie du changement est un modèle de la contribution aux effets attendus et non la cause de cet effet, car d’autres facteurs extérieurs peuvent aussi contribuer à l’obtention des résultats escomptés, comme l’indique la case « Influences extérieures ». La théorie du changement n’est un modèle de causalité que si aucune influence extérieure n’entre en jeu. Comme une intervention, une influence extérieure peut fonctionner isolément, mais elle peut aussi faire partie d’un autre faisceau de causalité qui pourrait inclure certains des facteurs de soutien du faisceau de causalité de l’intervention. Les influences extérieures peuvent avoir des effets positifs ou négatifs sur le niveau des résultats obtenus. Tout dépendant des forces de l’intervention, les influences extérieures peuvent expliquer tout, ou en partie, les résultats observés. Les effets négatifs importants, c’est-à-dire les risques susceptibles de saper la théorie du changement de l’intervention, sont inclus dans les hypothèses sur les liens de causalité.

En termes probabilistes, l’expression vraisemblablement suffisante peut servir à décrire le niveau de suffisance du faisceau de causalité de l’intervention, ce qui signifie que le faisceau de causalité a probablement contribué à l’obtention du résultat observé. Démontrer que l’intervention est une cause contributive équivaut à démontrer que son faisceau de causalité est vraisemblablement suffisant et que l’intervention, en soi, constitue un élément vraisemblablement nécessaire de ce faisceau.

Dans l’analyse des théories du changement, il est utile de distinguer le ex ante du ex post. Ex ante, il est nécessaire de parler des causes probabilistes et de la probable suffisance. Ex ante, on dispose d’une théorie du changement préliminaire ou postulée qui dit que si l’intervention est implantée telle que conçue et que si les hypothèses associées à la théorie du changement tiennent, alors la contribution anticipée aux résultats se concrétisera. Ce modèle assume la réalité et rend compte de la complexité de l’intervention. Il constitue une prédiction de l’efficacité de l’intervention.

Ex post, nous vérifions que la théorie du changement s’est effectivement réalisée, en utilisant les données probantes sur les résultats et sur la manifestation des hypothèses. Lorsque l’analyste conclut à une relation de causalité, il le fait en sachant quels facteurs étaient à l’œuvre et si d’autres éléments ont interféré. Si
l'analyste conclut que le faisceau de causalité était vraisemblablement suffisant, il reconnaît qu'un élément a peut-être échappé à l'analyse, mais qu'une personne raisonnable conclurait que le faisceau de causalité était effectivement suffisant. Ex post, nous vérifions l'hypothèse causale formulée ex ante. Comme il a été noté, ex post, nous sommes probablement en mesure de déterminer si d'autres influences extérieures sont entrées en jeu. Dans le cas contraire, nous pourrons dire que le faisceau de causalité de l'intervention a bel et bien causé le résultat observé, au-delà de la simple contribution.

L'intervention en soi est un des nombreux facteurs de causalité du faisceau de causalité nécessaire pour susciter un changement. En ce sens, tous sont égaux. Pourtant, notre intérêt se porte sur l'intervention comme instrument de changement, principalement sur les activités implantées dans le but d'obtenir un changement ou de le maintenir. Nous pouvons nous interroger, ex post, sur le rôle de l'intervention dans la survenu des changements. Nous pouvons nous attendre, au minimum, à ce que l'intervention serve de déclencheur pour démarrer la chaîne de causalité. Dans un tel cas, on pourra dire que l'intervention est une cause contributive principale. Parfois, l'intervention peut aussi être vue comme jouant un rôle complémentaire qui en s'ajoutant à d'autres éléments d'un processus déjà en cours, bonifie un changement déjà amorcé afin d'optimiser les résultats ou d'accélérer leur atteinte (Mayne, 2008).

**LES INTERVENTIONS MULTIFACETTES SUFFISANTES**

Malgré tout ce qu'elle contient, j'ai décrit la figure 1 comme une théorie du changement générique « de base ». C'est qu'elle ne présente qu'un seul acteur qui entreprend des activités; ce modèle peut suffire pour bon nombre d'interventions simples, mais pas pour la plupart des interventions complexes. Pour entraîner des changements, une intervention doit faire appel et s'associer à divers autres intermédiaires – partenaires de la prestation de services, administration publique, secteur privé, ONG – et influencer leur comportement. La théorie du changement illustrée à la figure 1 fait ressortir un éventail potentiellement large d'hypothèses de liens de causalité, qui doivent se concrétiser pour produire les avantages directs et les changements sur le bien-être. S'en remettre au hasard n'est peut-être pas une option, donc l'intervention doit s'associer à divers intermédiaires compétents, notamment des partenaires de la prestation de services, pour veiller à ce que les actions qu'ils entreprennent assurent (ou contribuent à assurer) la concrétisation des nombreuses hypothèses de liens de causalité. Ces activités de soutien menées par les acteurs de l’intervention s’ajoutent aux activités principales ou de base de l’intervention. On peut donc distinguer l’intervention de base et l’intervention globale. Dans une intervention sur la recherche en agriculture, l’intervention de base correspond aux activités de recherche, tandis que l’intervention globale regroupe ces activités plus celles qui soutiennent l’engagement et qui sont menées pour favoriser l'utilisation des résultats de la recherche. Dans d'autres cas, il n'y
a aucune activité de base ou principale : l’intervention s’associe à un éventail de partenaires pour offrir collectivement un ensemble d’activités suffisant.

Le plus souvent, ces hypothèses de liens de causalité couvrent une gamme d’événements ou de conditions qui créent un cadre supportant pour que les activités d’intervention contribuent au bien-être. Il en résulte une intervention globale multifacette, dont l’objectif est de rendre l’intervention suffisante : que l’ensemble des efforts d’intervention (de base), ses activités de mobilisation et les actions résultantes des intermédiaires influence effectivement les avantages et le bien-être. Autrement dit, l’ensemble des activités de mobilisation vise à assurer la concrétisation des hypothèses de liens de causalité, c’est-à-dire celle des facteurs de soutien.

Nous pouvons tout de même nous demander si l’intervention de base a été une cause contributive principale, autrement dit, si elle a contribué à déclencher le changement. Et dans les contextes de suffisance multifacettes, l’intervention implique la mobilisation de diverses actions de soutien afin de soutenir la réalisation du cheminement causal. Ainsi, il est important d’évaluer si l’intervention de base est une cause contributive de déclenchement et une cause contributive de soutien.

Une solide inférence causale au sujet d’une intervention multifacette suffisante serait que l’intervention a été une cause contributive principale des résultats pertinents observés :

L’intervention a été une composante nécessaire d’un faisceau de causalité qui, ensemble, ont été suffisants dans leur contribution au résultat observé. Autrement dit, l’intervention a fait une différence. En outre, l’intervention a joué un rôle clé : elle a déclenché et, par ses activités de soutien, appuyé la chaîne d’effets qui a contribué aux résultats observés.

La figure 3 illustre la théorie du changement générique dans cette intervention complexe, qu’on peut qualifier d’intervention multifacette8 suffisante.

Quand on construit une théorie du changement, il peut être utile de déterminer le degré de contrôle réel ou éventuel que l’intervention a sur les hypothèses de liens de causalité. On donne à chaque hypothèse une étiquette : [O], si l’intervention n’exerce aucune influence ou presque; [I], si l’intervention peut (devrait) exercer une influence directe ou indirecte; [C], si l’intervention devrait parvenir à exercer un contrôle direct. Cette stratégie aide à déterminer où des actions de soutien supplémentaires pourraient contribuer à la concrétisation des hypothèses et, partant, à une réduction au minimum des risques pour l’intervention, ce qui pourrait mener à une intervention multifacette.

À partir de l’exemple de la nutrition, le tableau 1 illustre le type d’analyse ex ante des liens de causalité que l’on peut entreprendre. Pour chacune des hypothèses de la figure 2, on évalue la mesure dans laquelle l’intervention pourrait investir des efforts pour renforcer la probabilité que l’hypothèse se réalise.

Dans l’exemple, il se peut que les maris ou les belles-mères soient peu susceptibles d’aider les mères à prendre des décisions sur la répartition des aliments. Donc, pour fonctionner, l’intervention nécessite une certaine forme d’éducation des maris et des belles-mères sur les avantages d’une alimentation nutritive pour
Figure 3. Théorie du changement de base pour les interventions suffisantes multifacettes
### Tableau 1. Analyse des hypothèses de liens de causalité dans une intervention en nutrition

<table>
<thead>
<tr>
<th>Hypothèse de lien de causalité</th>
<th>Degré de contrôle</th>
<th>Action de soutien nécessaire en plus des activités de base</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1 Hypothèses quant à la portée</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- La démarche et la documentation semblent adéquates.</td>
<td>Élevé [C]</td>
<td>Exige une bonne planification et une connaissance du contexte particulier.</td>
</tr>
<tr>
<td><strong>A2 Hypothèses quant aux changements dans les capacités</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Les bienfaits de la nutrition sont compris.</td>
<td>Élevé [C]</td>
<td>Exige une bonne planification et une connaissance du contexte particulier.</td>
</tr>
<tr>
<td>- Les pratiques alimentaires sont comprises et pertinentes.</td>
<td>Élevé [C]</td>
<td>Exige une bonne planification et une connaissance du contexte particulier.</td>
</tr>
<tr>
<td><strong>A3 Hypothèses quant aux changements de comportement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Les mères veulent améliorer la santé de leurs enfants.</td>
<td>s.o.</td>
<td>Peut être tenue pour acquise.</td>
</tr>
<tr>
<td>- Des aliments nutritifs sont disponibles à prix modique.</td>
<td>Élevé [C]</td>
<td></td>
</tr>
<tr>
<td><strong>A4 Hypothèses quant aux avantages directs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Les pratiques sont faciles à adopter.</td>
<td>Moyen [I]</td>
<td><strong>Action</strong> : Pourrait nécessiter un suivi pour vérifier si les pratiques sont faciles à adopter dans le contexte particulier.</td>
</tr>
<tr>
<td><strong>A5 Hypothèses quant aux changements sur le bien-être</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Les enfants ont accès aux services de santé.</td>
<td>?? [O]</td>
<td>Sera probablement simplement tenue pour acquise. Si la santé est un problème majeur, cela pourrait remettre l’intervention en question.</td>
</tr>
</tbody>
</table>
les enfants. L’organisme d’intervention devra peut-être amener d’autres parties prenantes qui ont davantage l’habitude de traiter des questions de culture et de genre, une ONG par exemple, à approcher les maris et les belles-mères afin d’influencer leur comportement. La figure 4 illustre l’intervention multifacette en nutrition qui en résulte.

L’analyse *ex ante* des liens de causalité peut aussi servir à évaluer *a priori* dans quelle mesure des données empiriques soutiennent chaque lien d’une théorie du changement. Parfois, il existe des recherches et des résultats d’évaluations antérieures qui permettent de documenter certains liens de causalité. Il arrive aussi que les informations existentes soient de faible qualité ou inexistantes. Dans ce dernier cas, il serait justifié d’entreprendre une nouvelle recherche afin de mieux confirmer le lien de causalité avant de mettre en œuvre l’initiative, ou encore d’assurer un suivi attentif des hypothèses rattachées à ces liens au cours de la mise

Bien sûr, l’analyse ex post des liens de causalité d’une théorie du changement est tout aussi intéressante. Déterminer dans quelle mesure un lien de causalité et ses hypothèses se sont réalisées et apprécier la crédibilité du cheminement causal est l’essence même de l’analyse de la contribution (Mayne, 2008; Mayne, 2012).

LES INTERVENTIONS COMPLEXES

Les figures 1 et 3 ciblent un seul groupe de bénéficiaires, les enfants dans l’exemple de la nutrition (figure 2). Toutefois, les interventions ont souvent plusieurs groupes cibles (les mères et les enfants, par exemple) ou plusieurs sous-groupes d’un groupe général (les garçons et les filles). Une intervention multifacette suffisante (figure 3) cible généralement plusieurs intermédiaires différents (administrations publiques, organismes, partenaires). Dans ce genre d’intervention à cibles multiples, une approche consiste à essayer d’élaborer une théorie du changement qui englobe toutes les activités menées sur les divers groupes cibles et les séquences de résultats qui s’ensuivent, de même que les liens entre les divers cheminement. Cependant, il risque de s’avérer très difficile d’élaborer et de représenter un tel modèle au-delà du simple survol (lequel a tout de même son utilité) et le modèle de théorie du changement complexe qui en résulte est souvent lourd et peu maniable, qu’il soit utilisé pour expliquer l’intervention ou pour concevoir l’évaluation.

Les théories du changement imbriquées

Il serait donc beaucoup plus utile d’élaborer, pour chacun des principaux groupes cibles, une sous-théorie du changement, appelée théorie du changement imbriquée ou théorie de la portée 9, chaque théorie étant éventuellement en interaction avec les autres théories dans la chaîne de production des effets. La figure 2 illustre les théories de la portée imbriquées (représentées par des ovales) pour les mères et pour les filles et les garçons. Garçons et filles pouvant faire l’objet de traitements différents, une théorie de la portée distincte pour chacun de ces sous-groupes ferait ressortir ces différences.

La figure 4 illustre la théorie du changement imbriquée applicable à une ONG dans l’exemple de la nutrition, qui est présenté en détail à la figure 5. Les hypothèses incluses dans la théorie du changement pour l’ONG (figure 5) sont celles que l’ONG devrait être à même de contrôler ou d’influencer fortement. La figure 6 illustre les théories de la portée imbriquées pour l’intervention multifacette suffisante générique. Les théories du changement imbriquées sont un moyen de décomposer une théorie du changement complexe en composantes plus maniables et plus faciles à comprendre.

L’analyse des liens de causalité

Pour traiter une théorie du changement complexe, on peut aussi analyser séparément les principaux liens de causalité, par exemple le lien entre « les changements
Figure 5. Théorie du changement imbriquée pour l’ONG
Figure 6. Théories du changement imbriquées
dans les capacités et les changements de comportement ». De plus, l'analyse des diverses hypothèses peut, par exemple, se faire dans un texte explicatif contenant des renvois aux recherches et évaluations antérieures qui appuient ces hypothèses, comme on l’a fait au tableau 1. Mayne et Johnson (2015) illustrent cette approche.

**L’incertitude et les résultats émergents**


**DIFFÉRENTES THÉORIES DU CHANGEMENT POUR DIFFÉRENTS OBJECTIFS**

Il y a des limites évidentes au niveau de détail qu’on peut inclure dans le modèle d’une théorie du changement, surtout dans le cas d’une vaste intervention multifacette. Nous avons besoin d’outils plus maniables, tant pour travailler le modèle qu’à des fins de communication. Il est donc utile de disposer d’au moins trois versions de chaque théorie du changement.

La première de ces versions, le *texte explicatif*, décrit en une ou deux phrases le fonctionnement prévu de l’intervention planifiée ou mise en œuvre. Cette version de la théorie du changement explique sans détour comment l’intervention est censée fonctionner; on peut aussi y énoncer les hypothèses sous-jacentes à l’intervention. C’est cette description de base, ce « synopsis », cet énoncé qu’un gestionnaire (ou un politicien) donne quand on lui demande d’expliquer pourquoi, à son avis, l’intervention va fonctionner. Prenons par exemple la théorie simple applicable à une intervention ayant la forme d’une publicité télévisée contre le tabagisme: si on décrit les dangers du tabagisme à la télé, des fumeurs cesseront de fumer. Dans l’exemple de la nutrition des enfants, la version « texte explicatif » de la théorie du changement pourrait se lire comme suit : « L’éducation et l’information des mères sur l’importance d’une alimentation nutritive pour leurs enfants les amèneront à modifier leur comportement et à chercher à améliorer le régime alimentaire de leurs enfants. » Ici, l’hypothèse sous-jacente sur laquelle repose l’intervention est qu’une meilleure information entraînera un changement de comportement.

Le texte explicatif de la théorie du changement joue un rôle important, car c’est ce message qui sera repris lorsque l’intervention sera présentée ou défendue.
sur la place publique. Ce message constitue, en quelque sorte, la version publique de la théorie du changement.

Une deuxième version de la théorie du changement, est un résumé de la théorie de changement, qui dresse globalement le portrait d’une intervention multifacette. Ce résumé peut être un schéma simplifié de la chaîne des effets indiquant les théories de changement imbriquées et les principales hypothèses. La figure 7 reprend l’exemple de la nutrition et de sa théorie du changement pour la sensibilisation des maris et des belles-mères représentée, ici, par un triangle.

La troisième version est plus détaillée. La théorie causale de changement est habituellement représentée sous forme de graphique, comme ceux des figures 1 à 5, qui indiquent la chaîne des effets et les différents liens de causalité. Chacune de ces versions de la théorie du changement a son utilité et il est souvent avantageux d’avoir les trois versions à sa disposition.

Un autre moyen de simplifier le modèle d’une théorie du changement consiste à omettre les « boîtes » et à inclure l’essentiel de leur contenu dans les hypothèses sur les liens de causalité, ce qui équivaut à réorganiser le faisceau de causalité. Pour bon nombre d’interventions, le regroupement de tous les éléments de la chaîne des effets ou de la théorie du changement dans un seul diagramme rend le graphique trop complexe et le lecteur se perd entre les flèches et les cases. La figure 8 illustre une version « simplifiée » de la figure 1, où les cases « Portée » et « Changements dans les capacités » ont été éliminées. Il est souvent tentant d’agir ainsi, car ce sont les changements de comportement qui sont considérés comme les principaux résultats du cheminement d’impact.

Comme nous le voyons dans la figure 8, la chaîne des effets est représentée de façon synthétique. Il faudra toutefois se rappeler que les aspects de la portée et des changements dans les capacités ne sont pas explicitement indiqués. Dans ce cas, les hypothèses allant des activités et produits de l’intervention aux changements

**Figure 7.** Survol d’une intervention en nutrition
de comportement devront inclure les hypothèses relatives à la portée et aux changements dans les capacités. Autrement dit, le faisceau de causalité associé à chaque lien demeure intact; il est simplement présenté différemment. L’omission de prendre en compte les enjeux associés à la portée et aux changements dans les capacités affaiblirait significativement la théorie du changement.

**LA CONSTRUCTION DES CHAINES D’EFFETS ET DES THEORIES DU CHANGEMENT**

Si concevoir et utiliser les théories du changement peut paraître compliqué, c’est uniquement parce qu’une « théorie du changement » n’est pas un objet en soi, un peu comme le concept d’« évaluation » peut représenter bien des choses différentes selon les situations. Les théories du changement :

- évoluent au fil du temps;
- ont des objectifs différents;
- doivent prendre en compte l’incertitude et la non-linéarité;
- peuvent être formulées *ex ante* et *ex post*.

www.theoryofchange.org traite en détail des théories du changement et propose des références à plusieurs autres guides et documents pertinents.

Pour obtenir une robuste première version d’une théorie du changement, il est nécessaire de tester la logique d’intervention et les hypothèses à la lumière de résultats de recherche ou d’évaluation existants qui pourraient confirmer (ou non) la théorie de changement établie. Brousselle et Champagne (2011) et Kauto et Silila (2005) discutent de cette mise à l’épreuve de la théorie de changement et suggèrent que cette étape est importante avant d’évaluer si l’intervention a donné des résultats. Ce type d’analyse permettrait d’identifier des points faibles dans les idées et hypothèses initiales sur le fonctionnement anticipé de l’intervention. Comme nous l’avons noté au départ, une théorie du changement peut aussi servir de cadre pour la conception de l’intervention, l’élaboration d’un système de monitorage et le développement d’un plan d’évaluation.

Lors de l’élaboration de chaînes d’effets et de théories du changement, il faut garder en tête plusieurs éléments :

- Il est probablement préférable de les établir de façon participative, bien que ce ne soit pas toujours possible.
- Dans un processus participatif, on peut soit commencer avec une page blanche et construire le modèle en sollicitant de différents points de vue, soit (ce qui est peut-être plus efficace) tracer d’abord, en petit comité, une ébauche de chaîne d’effets ou de théories du changement qui servira ensuite de base pour la discussion.
- Plus d’une version de la théorie du changement d’une intervention peut émerger des discussions avec les parties prenantes (Hansen et Vedung, 2010; Weiss, 1997). Il peut alors s’avérer utile de tester la vraisemblance de chacune des versions.
- Il est important d’inclure, de façon explicite ou implicite, tous les éléments de la théorie du changement.
- Le développement des chaînes d’effets et des théories du changement doit être considéré comme un processus; elles évoluent au fur et à mesure des connaissances acquises.
- Il faut accepter de produire une chaîne des effets ou une théorie du changement « acceptable » plutôt qu’une version parfaite.
- Les changements dans les capacités et de comportement sont souvent cruciaux dans les interventions.
- Autant que possible, les chaînes d’effets et les théories du changement devraient se fonder sur des recherches antérieures en plus du point de vue des parties prenantes.
- L’imbrication des chaînes d’effets et des théories du changement ou de la portée est un procédé très utile. Les éléments imbriqués s’élaborent autour de certaines stratégies ou groupes cibles de l’intervention.
- Les théories du changement peuvent être représentées de diverses façons et à divers degrés de détail.
• Des chaînes d'effets et des théories du changement génériques peuvent être très utiles pour représenter les principaux éléments d'une même intervention implantée dans différents contextes.

CONCLUSION
Établir des théories du changement crédibles est essentiel pour la conduite d'évaluations fondées sur la théorie (theory-based evaluations). Les modèles présentés ici se veulent assez adaptables pour s'appliquer à un large éventail d'interventions. Dans cet article, j'avance que les modèles de théorie du changement illustrés de façon générique aux figures 1 et 3 sont « utiles », pour plusieurs raisons :

• Ces modèles donnent souvent une représentation « acceptable » d'une théorie du changement sans verser dans l'excès de complexité. Ils jettent les bases pour l'élaboration d'histoires de la performance (Mayne, 2004).
• Ces modèles évitent délibérément d'utiliser la terminologie d'extrants, d'effets à court terme, d'effets intermédiaires et à long terme. Cette classification n'a pas de signification intrinsèque, n'est pas utile pour construire une théorie de changement et conduit trop souvent à des débats stériles. C'est davantage la séquence d'événement qui est importante. Les produits de l'intervention, la portée, les changements dans les capacités, de comportement, etc. tels qu'indiqués dans les figures 1 et 3 portent un contenu et offrent une meilleure structure analytique pour l'élaboration d'une théorie de changement.
• Les hypothèses sur les liens de causalité sont importantes pour décrire ce qui est nécessaire pour que le lien fonctionne; elles sont centrales dans la théorie du changement.
• Le recours à des boîtes pour représenter les hypothèses sur les liens de causalité simplifie la représentation et évite la prolifération des cases et des flèches.
• Il est souvent possible de simplifier un peu le modèle de théorie du changement en supprimant une case « résultat » et en incluant cet élément sous forme d'hypothèse.
• On peut simplifier les théories du changement en identifiant des théories imbriquées.
• Le modèle qui comprend les hypothèses sur les facteurs de soutien fait directement le lien entre le faisceau de causalité et les causes contributives, offrant une base rigoureuse pour une discussion sur la causalité.

NOTES
1 Le terme « modèle logique » est parfois utilisé comme synonyme de « théorie du programme » ou de « théorie du changement » (Funnell et Rogers, 2011), mais souvent, il
est une simple représentation de la chaîne causale. Ainsi, par exemple, le gouvernement du Canada définit « modèle logique » comme étant une « illustration de la relation de cause à effet, ou lien logique, entre les activités, les intrants, les extrants et les résultats d'une politique, d'un programme, ou d'une initiative, par exemple la chaîne des résultats » (Secrétariat du Conseil du Trésor, 2012).

Les principales corrections apportées visaient à simplifier la représentation du modèle en éliminant les références explicites aux « risques », aux « autres facteurs explicatifs » et aux « incitatifs » et à mentionner explicitement les résultats imprévisus.

Le savoir relève de l’information apprise ou des conseils suivis; les attitudes ont trait aux croyances, aux opinions, aux sentiments ou aux points de vue; les compétences renvoient aux habiletés mentales et physiques nécessaires à l’adoption de pratiques nouvelles ou de substitution; les aspirations englobent les ambitions, les espoirs, les objectifs ou les désirs. Adapté de Bennett et Rockwell (1995, p. 6).

Le « bien-être » est le terme générique utilisé ici pour décrire le résultat visé. On pourrait aussi parler du « mode de vie ».

Dans des articles précédents, j’avais souvent inclus explicitement les « risques » dans les cases des hypothèses et souligné que certaines hypothèses se comprenaient et s’énonçaient plus facilement sous forme de risques. Bien que parfois utile, cette technique a toutefois tendance à encombrer les cases.

Les théories de la portée imbriquées qui sont mentionnées dans la figure 2 seront abordées plus loin.

L’acronyme INNNS (INUS en anglais) représente une partie Insuffisante, mais Nécessaire d’une condition elle-même Non Nécessaire mais Suffisante pour que l’effet se produise (Mackie, 1974).

La figure 3 illustre un type d’intervention compliquée et complexe comportant des activités de base et de soutien. Un autre type d’intervention (non illustré), l’intervention multicomposante suffisante, se compose d’un certain nombre d’activités plutôt séparées et distinctes qui, mises ensemble, sont censées produire une amélioration du bien-être. Dans ce cas, les composantes seraient des théories du changement imbriquées dans une théorie du changement générale plus vaste.


Les points de la présente section sont développés plus avant dans Mayne et Johnson (2015).

**Références**


doi: 10.3138/cjpe.31144  
*CIPE* 32.2, 174–201 © 2017


**PRÉSENTATION D’AUTEUR**

Advancing Patient Engagement in Health Service Improvement: What Can the Evaluation Community Offer?

Nathalie Gilbert and J. Bradley Cousins
University of Ottawa

Abstract: Despite efforts for greater patient engagement in health care quality improvement, evaluation practice in this context remains mostly conventional and noncollaborative. Following an explication of this problem we discuss relevant theory and research on patient-centred care (PCC) and patient engagement and then consider potential benefits of collaborative and participatory approaches to evaluation of such initiatives. We argue that collaborative approaches to evaluation (CAE) are logically well-suited to the evaluation of PCC initiatives and then suggest contributions that the evaluation community can offer to help advance patient engagement. Finally, we outline a research agenda that identifies important areas that are in need of further examination.

Keywords: collaborative approaches to evaluation, health quality improvement, participatory evaluation, patient-centred care, patient engagement, program evaluation

Corresponding author: Nathalie Gilbert, University of Ottawa, Faculty of Education; nathalie.gilbert@sympatico.ca

© 2017 Canadian Journal of Program Evaluation / La Revue canadienne d'évaluation de programme
32.2 (Fall / automne ), 202–221 doi: 10.3138/cjpe.31120
STATEMENT OF THE PROBLEM

Health organizations are increasingly making efforts to offer patient-centred health services that are more responsive to patients’ preferences, values, and needs. The underlying philosophy of patient-centred care (PCC) advocates for patients to have an active role in all areas of their care, including broader areas of the health care system such as planning and evaluation of health services (Gerteis, Edgman-Levitan, Daley, & Delbanco, 1993; Stewart et al., 2000). The conceptualization of PCC emerged during the 1950s and is widely used in health organizations to refer to an essential component of quality health care. The concept of PCC was originally introduced as a rejection of the traditional model of care that tended to be more disease- and physician-focused, toward a health care that is more holistic in nature and aims to improve the patient experience. Studies have demonstrated that PCC, with its emphasis on being responsive to patient preferences and needs, has positive outcomes on the quality of health care and is strongly desired by patients and families (Agency for Health Care Research and Quality, 2005; Little et al., 2001; Rao, Weinberger, & Kroenke, 2000).

Current strategies used for the assessment and monitoring of PCC and patient experience tend to be primarily based on performance measurement and patient/family feedback (Baker, 2014; Davies & Cleary, 2005). For example, patient satisfaction surveys are commonly used for reporting on the quality of health services as well as to inform decision-making. Considered one of the gold standards in health care quality improvement, patient satisfaction surveys have a wide range of validated surveys available that are low-cost, can gain large amounts of feedback, can be easily implemented, and can provide hospitals with an opportunity to receive posthoc feedback from patients (Canadian Foundation for Healthcare Improvement [CFHI], 2012; Veillard et al., 2005). Yet, as reported by the CFHI, there is a myth in health care that “high patient satisfaction means high quality care” (2012, p. 1). Martin and Ronson (2007) caution that an approach in which health organizations rely on patient satisfaction surveys to identify areas that need improvement and to learn more about the patient experience is insufficient. They concluded that “fifty years of patient satisfaction research has found that in study after study between 80 and 90% of patients are satisfied” (2007, p. 8). Some of this could be explained by the fact that most patient satisfaction surveys tend to ask surface-level questions (i.e., discrete, categorical questions about delivery of care), are subject to self-selection and literacy bias, and are therefore limited in capturing the patient’s experience beyond the questions being asked (CFHI, 2012; Williams, Coyle, & Healy, 1998). Other forms of feedback also commonly used in tapping into patient views about patient experience include interviews, complaints, and patient or family narratives. They are beneficial in capturing patient experience information, but tend to focus on individual issues, are resource intensive, and often challenge decision makers on how to incorporate these data into quality improvement plans (Baker, 2014).

Another common practice in health care quality improvement efforts is the use of performance measurement systems, which monitor the performance
of the broader health care system using statistical data to determine progress toward specific defined objectives (Adair et al., 2006). Health decision makers use performance monitoring data by monitoring selected indicators and targets to help assess their performance over time and to make comparisons with other health institutions that offer similar services. Performance data serve an important role in ensuring health organizations meet their accountability requirements established by their funding agencies. They are also regularly used to help inform health professionals and decision makers on how health consumers use the health system (e.g., number of emergency department visits, wait times) as well as for establishing strategic directions. However, an important limitation is the relative lack of health performance systems that include indicators that can further our understanding of patients’ health service experiences, integral to improving PCC (Canadian Institute for Health Information [CIHI], 2016). One explanation for this may be that patients or family members are not regularly involved or consulted in the development of health services indicators. That is, performance indicators are typically developed through provincial or state requirements, research evidence, and consensus of an expert panel, usually comprised of health professionals without input from patients or family members (Baker, Fancott, Judd, & O’Connor, 2016). As a result, performance data often fall short in terms of informing decision makers about issues that are of importance to patients and families. For example, administrators, clinicians, and patients give importance to different aspects of care, and their ideas on quality health services differ (Kötter, Schaefer, Scherer, & Blozik, 2013). Administrators tend to be concerned with operational efficiency (e.g., bed occupancy, budget), clinicians are interested in clinical efficiency (e.g., mortality rates, complications) and patients tend to be more concerned with the health care experience and interpersonal interactions with health providers (Ioan, Nestian, & Tiță, 2012). Arguments have been made for greater involvement of patients and family members in indicator development and selection to collect performance data that is more reflective of patient and family priorities and to better understand the patient experience (Gagliardi, Lemieux-Charles, Brown, Sullivan, & Goel, 2008a, 2008b; Kötter et al., 2013). Recently, the CIHI has taken the lead in making efforts to address the indicator gaps related to measuring patient experience and are in the process of developing “a set of indicators to measure patient experience, inform performance improvements over time, and support benchmarking across Canada” (CIHI, 2016, p. 2). This is an important contribution to health performance systems and will certainly advance health professional and decision makers’ abilities to consider and better understand patient experiences. However, key challenges that will remain are the ability to gain a true understanding of the local context of patient experience as well as explaining the variance in results.

The use of performance measurement and patient feedback is well entrenched within the health sector and highly valued by decision makers. These strategies serve important functions in terms of meeting accountability requirements and monitoring patient satisfaction. Yet, despite efforts to collect these forms of data,
the application of quality improvement strategies that would facilitate patient-centredness across organizations remains a challenge and has had mixed results (Baker, 2014; International Alliance of Patients’ Organizations [IAPO], 2007; Ponte et al., 2003; Robinson, Callister, Berry, & Dearing, 2008). For example, Baker (2014) reviewed how patient engagement contributes to improved care, reporting that decision makers have reported feeling challenged in linking these data sources to answer key priority questions. Overall, efforts made toward developing sophisticated measures for capturing patient experience have not led to improved knowledge on how to apply these results as a means to improving patient experience. Baker (2014) noted, “While data and stories about patients are important sources of information, they may be insufficient to motivate and focus improvement in many contexts” (p. 2).

The Institute of Medicine report, *Crossing the Quality Chasm* (IOM, 2001), has been influential and has received widespread support within health care for applying more patient-focused quality improvement strategies. These strategic efforts toward increasing patient engagement are geared toward addressing some of the gaps identified in the use of conventional quality improvement strategies as well as to improve the quality of patient-centred health services. More recently, there has been a sharp increase in the number of government and nongovernmental organizations sponsoring and advocating for greater patient engagement activities across all levels of health care (e.g., quality improvement, accreditation, strategic planning, research, etc.). Two of many examples are the *Patients First: Action Plan for Health Care* that made a commitment “to expand patient engagement” in Ontario (Ministry of Health and Long-Term Care, 2015) and *Canada’s Strategy for Patient-Oriented Research*, providing a “continuum of research that engages patients as partners, focusses on patient-identified priorities and improves patient outcomes. This research, conducted by multidisciplinary teams in partnership with relevant stakeholders, aims to apply the knowledge generated to improve healthcare systems and practices” (Canadian Institutes of Health Research, 2016, para 2).

To date there is limited empirical research that has examined the effects of patient engagement or the best approach to engage patients (Abelson et al., 2015; Baker, 2014). Furthermore, there is a relative lack of collaboration and shared knowledge between the evaluation community and health sector in the rapidly developing area of patient engagement and the development of best practices. As a consequence, health organizations continue to struggle on how best to involve patients (i.e., process) in health service improvement initiatives as well as learn from patient experience (Baker, 2014; Luxford, Safran, & Delbanco, 2011). In this article we make the argument that the evaluation community could offer significant contributions in these areas through the promotion of evaluation approaches that are collaborative and participatory—approaches that are logically and conceptually aligned with PCC initiatives.

With greater attention and efforts being made toward patient engagement in the planning, design, and evaluation of health services, it is critical to identify
strategies to continue the advancement of patient engagement as complementary to improving patient-centred health services. Health organizations have made efforts to increase patient engagement efforts, but some of these developments have not always been as productive and effective as originally intended. In this article, we discuss the relevant literatures of PCC and patient engagement to further our understanding of some of the challenges and facilitators that might allow patient engagement to prosper. Following this discussion, we offer the use of collaborative approaches to evaluation (CAE), such as practical participatory evaluation (P-PE), as potentially powerful facilitators of patient engagement that are currently missing. Through collaborative practices we believe the evaluation community can help advance patient engagement in health care quality. Finally, we outline a research agenda that identifies several important areas that are in need of further examination. We begin by providing an overview of the evolutionary trajectory of PCC, providing a conceptual perspective for discussing some of the facilitators, barriers, and approaches to patient engagement.

**PATIENT-CENTRED CARE**

The conceptualization of PCC emerged during a time of substantive change in health care, changes that saw the introduction of new technologies, specialties, and patient populations. PCC is entrenched in health organizations to refer to an essential component of quality health care. Despite difficulties defining the term, by the late 1990s there were some general principles associated with PCC, many of which were based on the seminal work produced by the *Picker Commonwealth Program for Patient-Centred Care*, now known as the Picker Institute, which is considered to be one of the more influential organizations for advancing PCC (Shaller, 2007). The Picker inquiry, led by Gerteis et al. (1993), conducted a wide range of focus groups of recently discharged patients, family members, physicians, and nonphysician hospital staff as well as reviewed pertinent literature related to PCC. Seven key principles of PCC were identified through this work, which culminated in the publication of a book entitled *Through the Patient’s Eyes* (Gerteis et al., 1993). The fundamental PCC principles identified in their work were

- respect for patients’ values, preferences, and expressed needs;
- coordination and integration of care;
- information, communication, and education;
- physical comfort;
- emotional support and alleviation of fear and anxiety;
- involvement of family and friends; and
- transition and continuity. (p. 223)

Subsequent to the publication of this book, there was a significant increase in the interest in PCC across health care systems and significant uptake in the efforts to define PCC (IAPO, 2007). For example, in its landmark report, *Crossing the*
Quality Chasm, the IOM (2001) included PCC as one of their six aims for health care improvement and declared that “making care more patient-centred means adjusting nearly every aspect of practice, in every realm from the administrative to the clinical to the technological” (p. 1).

Despite the prominence given to PCC within the health care system, implementation of effective PCC practices remains a challenge (Luxford et al., 2011; Shaller, 2007). Identification and distinction of PCC definitions within the different health care contexts in which it is being applied could assist health care providers in improving the implementation of PCC. For example, Robinson et al. (2008) found that most PCC definitions are derived from four distinct sources: a public policy perspective, an economic perspective, a clinical perspective, and a patient perspective. The PCC perspectives that Robinson et al. (2008) have suggested deconstruct the PCC concept into more manageable and context-specific components. These four proposed PCC perspectives challenge health organizations to focus their PCC efforts at all levels of the health care system and organization. Significant efforts have been made in the past decade, using a bottom-up approach, to improving PCC practices within the patient and clinical PCC perspectives that take place at the individual patient level. However, one area of development that deserves greater attention is exploring PCC practices aimed at the broader, organizational level of health programs and organizations (IAPO, 2007). Robinson et al. (2008) refer to this as the public policy perspective of PCC, and it serves as the foundation for all other PCC practices and sets the direction for PCC across health care. Patient engagement in setting research priorities or in the planning and evaluation of health services are examples of PCC being implemented at the organizational level, which is the focus of this article.

Despite considerable efforts in applying quality improvement efforts (e.g., incident reporting, clinical audit, risk management, etc.), health organizations have had mixed results in implementing widespread patient-centredness (Luxford et al., 2011; Shaller, 2007). However, further research has examined critical organizational facilitators to promoting successful implementation of PCC. The most critical include (a) a strong and committed leadership (i.e., CEO and board of directors), identified as critical for achieving sustained delivery of PCC; (b) a strong and clearly communicated strategic vision that was constantly communicated to every member of the organization; (c) systematic measurement and regular feedback to health service providers (e.g., front-line staff, decision makers, etc.) of patient experience data with high specificity; (d) involvement of patients and families at multiple levels (e.g., service redesign, partners in care, patient and family advisories, representation on medical executive committees, etc.); and (e) a culture that strongly supports change and collective learning, identified as a powerful enabler to PCC (Luxford et al., 2011).

While the concept of PCC is well entrenched in health care, there continue to be challenges in its meaning and implementation. The factors contributing to successful implementation of PCC outlined in the previous section may provide some insight into practices that elevate the norm. In response to the continuing
The challenge of incorporating widespread patient-centredness across health organizations, momentum is building and there is commitment within the health sector to seriously support activities that engage patients in health care quality improvement efforts.

**PATIENT ENGAGEMENT**

The following review of the relevant patient engagement literature is based on material acquired through systematic searches of the knowledge base using standard computerized search and retrieval tools. Specifically, we sought empirical and conceptual published research in comprehensive health management databases using Medline (Ovid), Allied and Complementary Medicine, Embase, Healthstar, PsycINFO, and Cochrane Database of Systematic Reviews using the keywords “patient engagement” and “patient involvement.” Sources addressed the following questions: (a) How is patient engagement defined/described for health care quality improvement? (b) What are the facilitators and challenges to patient engagement in planning and evaluation of health services? (c) What does the approach or process used for engagement look like? The search was limited to the years 2000 to present and targeted peer-reviewed published work. In addition to the computerized search, we used bibliographic follow-up to identify relevant publications.

This led us to a body of knowledge that promotes *patient engagement* in planning and evaluation as a means of addressing PCC at the broader level of a health program or organization. The terminology surrounding patient engagement, patient involvement, public engagement, and public involvement in health care is quite varied. All of these terms share the common goal of seeking public or patient input and guide health system decision-making on specific health care issues (Abelson et al., 2016). The term *public engagement* appears to be used as an umbrella term in the literature for public involvement in health decision-making; other terms such as *patient engagement/involvement* tend to have more limited specific meanings and different intensities of engagement (Hill, O’Grady, Millar, & Boswell, 2000; Mitton, Smith, Peacock, Evoy, & Abelson, 2009). Abelson et al. (2015, p. 2) describe *public and patient engagement* as a “term to capture a wide range of efforts aimed at actively involving citizens and patients in various domains and stages of health system decision-making.” This article is only concerned with the patient perspective rather than the broader societal public perspective; therefore, the term *patient engagement* will be used to represent the active engagement of patients, including family members, in health service planning and evaluation. The CFHI has been at the forefront of supporting patient engagement initiatives and describes patient engagement as

initiatives that engage patients and families in designing, delivering, and evaluating health services, with the goal of improving the quality of care. Co-designing improvements with patients and families leads to new insights and better results than providers and leaders working on their own. (CFHI, 2017)
Patient engagement gained considerable attention in North America and Europe more than a decade ago, and the practice of patient engagement for health service improvement has evolved considerably in Canada (Abelson et al., 2015; CFHI, 2017; Crawford et al., 2002; Van de Bovenkamp, Trappenburg, & Grit, 2010). The United Kingdom and the Netherlands have been at the forefront of developing strategies for patient engagement (e.g., experience-based co-design, Bate & Robert, 2007) as well as conducting empirical studies examining the effects of patient engagement (Crawford et al., 2002; Fudge, Wolfe, & McKevitt, 2007; Van de Bovenkamp et al., 2010). In Canada, efforts toward greater patient and family engagement for improving patient experience have been more recent. For example, the Change Foundation, an independent policy think tank, added a strategic priority that “will focus on listening and learning to better understand the family caregiver experience as part of the patient experience and to identify promising models or initiatives for effective and collaborative engagement between family caregivers and providers” to their 2015–2020 strategic plan (Change Foundation, 2015 p. 16). Cancer Care Ontario’s 2011–2015 Ontario Cancer Plan identified “continue to assess and improve the patient experience” as one of their six strategic priorities (Cancer Care Ontario, 2011, p. 40). Efforts toward increased patient engagement in quality improvement have become priorities for many health organizations as a means of improving patient experience, but progress has been limited (Baker, 2014; Baker et al., 2016).

Carman et al. (2013) have made significant contributions in defining and describing what is involved with patient engagement and proposed a multidimensional framework for patient and family engagement in health and health care. The framework describes the levels at which patient engagement can occur across the health care system, from direct care to patient engagement into organizational design, evaluation, governance, and policy-making. The framework depicts how much information is exchanged between patient and provider as well as how active a role the patient has along the continuum of engagement. For example, at the continuum’s lower end, which tends to be consultative in nature, “patients are involved but have limited power or decision-making authority. Providers, organizations, and systems define their own agendas and then seek patients’ input. Information flows to patients and then back to the system” (Carman et al., 2013, p. 224). At the higher end of the patient engagement continuum, “engagement is characterized by shared power and responsibility, with patients being active partners in defining agendas and making decisions. Information flows bidirectionally throughout the process of engagement, and decision-making responsibility is shared” (Carman et al., 2013, p. 224).

There is limited empirical research that has examined the process, effects, and best approaches for engaging patients in health organizational design/governance, evaluation, and policy (Armstrong, Herbert, Aveling, Dixon-Woods, & Martin, 2013; Baker, 2014; Baker et al., 2016). The following section provides an overview of the literature on the facilitators and challenges of patient engagement in the planning and evaluation of health services that go beyond direct care.
Strengths and Facilitators of Patient Engagement

Baker and Denis (2011) identified patient engagement initiatives as a means to better respond to patients’ needs and expectations as well as a priority area for system change within Canada. They also suggested that patient engagement is increasingly being seen as a “potentially strong lever to shift the system toward improvement and to align the perspectives and activities of different practitioners” (p. 25).

Review of patient engagement studies indicates that there has been a positive impact in the following areas: improved educational or tool development (Baker et al., 2016), possibly improved clinical care outcomes or service delivery (Baker, 2014; Baker et al., 2016; Rathert, Wyrwich, & Boren, 2013), informed policy or planning initiatives as well as accelerated work and its visibility (Baker, 2014), and improvement in patient experience (Baker et al., 2016). Some facilitators identified in the literature to enhance patient engagement or patient-centred care initiatives include (a) having a clear rationale for patient engagement and identifying the right patient engagement approach to achieve the desired outcomes, (b) dedicated champion and/or committed leadership that communicates strategic vision across the organization, (c) building staff capacity to work with patients, (d) having adequate resources, (e) organizational culture that is committed to change, (f) learning and involving patients in a meaningful way, and (g) clear roles and responsibilities (Armstrong et al., 2013; Baker, 2014; Baker et al., 2016; Luxford et al., 2011; Shaller & Darby, 2009). Crawford et al. (2002) reported that staff attitudes toward collaborating with patients became more favourable compared to a baseline and that the organizational culture became more open to working with patients as a result.

Recent development of a public and patient engagement evaluation tool developed by Abelson et al. (2015) could be an important facilitator in advancing patient engagement and evaluating its impact. The tool was developed through a collaborative process that involved review of the literature as well as input from public and patient engagement researchers and practitioners. The four principles identified as critical for the evaluation of public and patient engagement are as follows: (a) integrity of design and process (i.e., diverse range of views, clear communication between organizers and participants, and support to enable participation), (b) influence and impact (i.e., informs planning/decision-making, learning, and increased confidence and trust), (c) participatory culture (i.e., organizational support for patient engagement, leaders informed on patient engagement, and demonstrated use of patient engagement work), and (d) collaboration and common purpose (i.e., plan and coordinate collaboratively to address concerns of people they serve). Despite the strengths of patient engagement in health service planning and evaluation, it does not come without its unique challenges. This next section will discuss some of the challenges faced when engaging patients in planning and evaluation initiatives.
**Challenges to Patient Engagement**

The vagueness of the concept of *patient engagement* and the lack of clarity of purpose, structures, and roles among patients and health professionals involved in a patient engagement initiative were reported to challenge the patient engagement process (Armstrong et al., 2013; Baker, 2014; Crawford et al., 2002; Fudge et al., 2007; Gagliardi et al., 2008a; Tedford-Gold, Abelson, & Charles, 2005). Some view patient engagement as a quality issue that can contribute to improvements of services, but others see it as simply an accountability requirement that can be accomplished with a single time point consultation rather than genuine engagement.

Studies examining patient experiences with patient engagement were mixed. Some patients reported that the engagement process was rewarding and appreciated the opportunity to be involved; other studies reported patient dissatisfaction with the process and lack of interest in being involved in health improvement initiatives (Fudge et al., 2007; Gagliardi et al., 2008a). There was also some reluctance on behalf of some health professionals to support shared decision-making with patients (Frosch, May, Rendle, Tietbohl, & Elwyn, 2012). The variations in the valuing of patients’ experiential knowledge resulted in some studies reporting a significant gap between their intentions to involve patients and what actually occurred. Finally, appropriate guidelines or approaches for engaging patients in health service planning and evaluation are quite lean. Participatory evaluation and research approaches to patient engagement have been suggested as a good starting point, but there is little published evidence that such approaches have been used (Armstrong et al., 2013; Bate & Robert, 2007).

Methodologically, few studies examined the effectiveness or quality of their patient engagement approaches. Most studies examining patient engagement approaches were descriptive in nature and were generally restricted to authors’ reflections on the strengths and limitations of their engagement process. To date, most patient engagement in program improvement initiatives have been consultative in nature rather than representing genuine interactive engagement sustained over time. Very few studies examined the patients’ experiences or the impact of the engagement process on patients and health professionals involved in the patient engagement process; considering that the intent of patient engagement is to be more patient-centred, this is somewhat ironic (Fudge et al., 2007; Gagliardi et al., 2008a). Also of concern is the relative lack of collaboration and shared knowledge between the evaluation community and health sector related to patient engagement. Many of the strengths and challenges identified and discussed in the patient engagement literature have a long history in the evaluation literature. Evaluators have unique skills and evaluation approaches that could make significant contributions in ensuring that patient engagement prospers and ultimately improves the patient experience. In the next section, we provide an overview of collaborative approaches to evaluation and discuss how such approaches could serve the improvement of patient engagement processes and strengthening partnerships between patients and health service providers.
COLLABORATIVE APPROACHES TO EVALUATION AND PATIENT ENGAGEMENT

The very nature of patient engagement in health service planning and evaluation implicates the use of collaborative and participatory approaches. Yet one of the barriers to patient engagement in health service planning and evaluation is the limited knowledge on how to put such approaches into practice (Fudge et al., 2007). Recent developments in the evaluation community, particularly with regard to the development and validation of principles for CAE (Shulha, Whitmore, Cousins, Gilbert, & Al Hudib, 2016) offer useful guidance. CAE seeks to develop a partnership between the evaluator and members of the program or intervention community such as program developers and managers or service users (e.g., patients). Evaluators bring technical expertise and knowledge of evaluation professional standards of practice to the planning and evaluation process, and patients bring experiential knowledge with the interventions (i.e., health services) and a rich knowledge of the context in which the interventions are implemented (Cousins & Chouinard, 2012; Cousins & Earl, 1995; Stevahn, King, Ghere, & Minnema, 2005). A key element to CAE is that stakeholders (e.g., patients, family members, health professionals, decision makers) are actively involved in decision-making and are able to see tangible evidence of their contributions. Of equal importance in CAE is attention to capacity-building designed to assist program community members to understand the evaluation process and their role in it. Cousins and Earl (1995) and Cousins and Chouinard (2012) view participatory approaches to evaluation as a means of enhancing the use of evaluation findings and working toward creating an organizational culture that is committed to learning and improvement. Learning not only involves quantitative standards from performance indicators, but from genuine collaboration between health practitioners/researchers and patients/family or process use (Patton, 1997).

Figure 1 elucidates the eight principles to guide CAE that were recently developed and validated by Shulha et al. (2016). A growing number of participatory approaches fall under the CAE umbrella (e.g., practical participatory evaluation, most significant change technique, rapid rural appraisal). Common to them all is that evaluators work in partnership with members of the program community. Three important considerations are associated with these principles: first, they are to be considered as a set, not as a pick-and-choose menu for application; second, they are well differentiated from yet overlapping and interconnected with one another; finally, they are not intended to imply a linear sequence, although there is a loose temporal order beginning with “Clarify the Motivation for Collaboration.”

The principles can be used most importantly as a guide to CAE practice but also to retrospectively analyze projects, review evaluation policy, and inform professional development, among other applications. In the present case, the emphasis placed on the “participants’ central role in the evaluation process is of great interest since it parallels Mallett’s (1996) definition of PCC, which emphasizes placing patients in the centre of the system of care. It is important to note that many types of CAE involve other stakeholders such as managers and implementers in evaluation.
Figure 1. An Integrated Set of Principles for Use in Guiding Collaborative Approaches to Evaluation

Patient engagement provides patients with an opportunity to dialogue with health professionals about what is of importance to them as well as to be actively involved in decisions for improving health services, a key principle of PCC. However, recurring weaknesses from the patient engagement literature are the lack of guidance or approach to implementing patient engagement into practice, the gap between intentions to involve patients and their actual involvement, patient engagement tending to be more consultative than collaborative, and the challenges and time required to shift clinician attitudes to adopt a more “patient and family focus” to their practice (Baker, 2014; Fudge et al., 2007; Tedford-Gold et al., 2005). For example, in Gagliardi et al.’s (2008a) study, one of the suggested recommendations made by participants was that the patient engagement “process should be ongoing and interactive rather than single, passive efforts to enable information sharing, and foster mutual understandings of perspective among patients and health professionals” (p. 239). This type of recommendation bodes well for a patient engagement approach that is collaborative in nature, similar to CAE.

In considering justifications for CAE, Cousins and Whitmore (1998) identified two streams of participatory evaluation: (a) Practical Participatory Evaluation (P-PE) and (b) Transformative Participatory Evaluation (T-PE). P-PE supports program or organizational decision-making and its main function is in fostering
evaluation use. “The core premise of P-PE is that stakeholder participation in evaluation will enhance evaluation relevance, ownership, and thus utilization” (p. 6). The second rationale, T-PE, seeks to empower members of community groups who are less powerful and “invokes participatory principles and actions in order to democratize social change; it has quite different ideological and historical roots from P-PE” (p. 7). For the purpose of this article, the proposed use of participatory evaluation approaches would fall within the P-PE rationale. Despite there being an element of empowerment in patient engagement, the overall goal is to improve the quality of care while recognizing that the process of patient engagement has a very practical aspect to it. In addition, it promotes utilization of evaluation results by having decision makers involved throughout the evaluation process. It should be noted, however, that capacity building and empowerment outcomes may also occur in P-PE and that transformative and practical outcomes are by no means mutually exclusive.

Three fundamental dimensions of process in CAE have been identified by Cousins and Whitmore (1998) and affirmed by others (e.g., Daigneault & Jacob, 2009). These dimensions, appearing in Figure 2, are considered to be orthogonal and represent decision points that help to shape the CAE process. The first is control of evaluation process—who controls the decision-making related to the technical aspects of the planning and evaluation process: researchers/evaluators, organizational decision makers (health professionals)/service users (patients), or some balance between the two? The second dimension is stakeholder diversity—who (e.g., program decision makers, patients, family, program staff) within the program or organization should be involved in the planning and evaluation process? The final dimension is depth of participation, implicating the intensity of involvement in the evaluation process, ranging from light touch consultation to significant engagement with all phases of evaluation planning and implementation.

We argue that widespread use of CAE in the context of PCC is lacking, but it has enormous potential to assist health organizations to leverage patient engagement in meaningful and sustainable ways. We advocate practical CAE approaches to broaden decision-making and problem-solving by engaging a range of key stakeholders in planning and conducting evaluation for health service quality improvement initiatives. Intended service beneficiaries (patients) would be the central figures among participating stakeholders, given compatibility with the principles of PCC. The direct use of the CAE principles to guide planning and evaluation efforts has considerable potential to assist health organizations to further their agenda of fostering patient engagement.

CONCLUSION AND AGENDA FOR FUTURE RESEARCH

Our survey of the patient engagement landscape, while likely not exhaustive, provides an overview of current quality improvement methods being used in the health sector as well as some of the complexities and challenges of engaging patients and health professionals in collaborative program improvement initiatives.
Figure 2. Dimensions of Form in Collaborative Inquiry (Adapted from Cousins and Chouinard, 2012)
The complexity of engaging patients in planning and evaluation requires continued research that goes beyond examining patient engagement at the level of direct care, toward a better understanding of patient engagement at a program or system level. Despite emerging evidence that suggests patient engagement leads to patient and organizational improvements (Baker, 2014), we need to explore and gain a better understanding of the processes involved or evaluation approaches that could contribute to translating patient engagement into improved outcomes. In this final section, we identify priority issues and questions for study that we believe could help further our understanding of patient engagement in health care improvement.

1. **Practical CAE**: Despite significant effort, the extent to which patients have been engaged in health care quality improvement has mostly been limited to various forms of single time point consultation, rather than a genuine interactive partnership based on principles of CAE. Of interest would be the examination of practical forms of CAE for the development of effective engagement processes. Such activities carry significant potential to leverage evidence-based decision-making in the interest of service improvement.

2. **Facilitators and barriers**: In our review, few patient engagement initiatives evaluated or examined the patient’s or health professional's experience with the engagement experience. How do patients feel about their engagement experience? What are some of the factors that facilitate or restrain patient engagement? To what extent does more intensive involvement in evaluation than is presently the case mediate such factors?

3. **Process considerations**: Do we know what approach is best suited to engage patients in the evaluation process? Principles for CAE implicate the development of deep understandings of context and the nature of interventions to inform subsequent process decisions such as (a) who controls the technical evaluation decision-making (evaluator, balanced, stakeholder)? (b) who from the health care community (e.g., managers, caregivers), apart from patients, should be involved in evaluation and why? (c) how intensive (light touch, deep involvement) should patient and other stakeholder engagement in evaluation be?

4. **Understanding consequences**: It is important to understand patient engagement consequences in the health care setting context. What are the observed effects of the engagement process? Has improved patient experience or patient centredness been achieved? What are the intended benefits of engaging patients in the planning and evaluation process? What are the unintended effects (positive or negative) of engaging patients in planning and evaluation processes? To what extent did these observed outcomes depend on process dynamics?

5. **Capturing perspectives**: We are of the view that alternative methodological approaches are required for assessing patient experience and the
effects of patient engagement beyond measurement and patient feedback (i.e., patient satisfaction surveys, patient narratives). Performance data often fall short in informing decisions about issues that are of importance to patients, and patient satisfaction surveys are limited in terms of identifying areas that need improvement or learning about the patient experience. We would suggest complementing these strategies with ones that provide deeper understanding of the patient experience and help focus improvement efforts—for example, qualitative approaches, such as participant observation, as a means of giving primacy to what is meaningful to patients and focus improvements. In short, move beyond the reflective narrative in describing the patient engagement process and its impacts on PCC.

We have provided an overview of patient engagement within a health care context and identified important areas where we believe the evaluation community can play a significant role in continuing to advance the body of knowledge and practice of patient engagement. Our intention has been to explicate some options for more collaborative practice and to inform a future research agenda that will add new insights and advance our understanding of patient engagement. In the end, our hope is that research on patient engagement will help highlight the importance of including the patient perspectives to create a more responsive health care system.

ACKNOWLEDGEMENTS

The authors acknowledge financial support for this research provided by the University of Ottawa. The opinions expressed herein do not necessarily reflect those of the sponsoring agency.

REFERENCES


**AUTHOR INFORMATION**

Nathalie Gilbert is a Ph.D. candidate in the Faculty of Education, University of Ottawa. Her research interests are in collaborative approaches to evaluation, and she is currently conducting research examining the process and impacts of engaging patients and their families in the planning and evaluation of health services.

J. Bradley Cousins is Professor Emeritus of Evaluation at the Faculty of Education, University of Ottawa. Cousins’ main interests in program evaluation include participatory and collaborative approaches, use, and capacity building. Throughout his career he has published many articles and books on evaluation, received several awards for his work and served as the editor of the *Canadian Journal of Program Evaluation* from 2002 to 2010.
L’Ombudsman face aux défis de l’évaluation : est-il possible d’évaluer l’intangible?

Seyive Wilfried Affodegon, Steve Jacob et Éric Montigny
Université Laval

Résumé : L’évaluation de la performance de l’Ombudsman est complexe en raison d’une mission de nature procédurale et de caractéristiques spécifiques aux organismes de protection des droits. Cet article offre une synthèse des connaissances de la mesure de sa performance et ouvre la voie à de nouvelles réflexions tant théoriques que pratiques. En plus de présenter les forces et les faiblesses des différents modèles d’évaluation recensés dans la littérature, cette étude précise la raison d’être de l’Ombudsman ainsi que les finalités de son évaluation. Cet article formule les prémisses d’une évaluation intégrée de la gestion de l’Ombudsman parlementaire.

Mots clés : critères d’évaluation, ombudsman, protecteur du citoyen

Abstract: Owing to its procedural mission and the specific characteristics of rights protection organizations, the evaluation of an Ombudsman’s performance is a complex task. This article presents a knowledge synthesis of performance measurement for Ombudsman’s Offices and opens the way to new reflections that are both theoretical and practical. In addition to presenting the strengths and weaknesses of the various evaluation models identified in the literature, this study clarifies the purpose of the Ombudsman and the purposes for which it is evaluated. Finally, this article formulates propositions for an integrated assessment of the management of the Parliamentary Ombudsman.

Keywords: assessment criteria, Ombudsman

INTRODUCTION
Les nouvelles approches de gestion publique accordent un intérêt particulier à la mesure de la performance et à l’évaluation des organismes publics (Behn, 1995; Heinrich, 2002; Pollitt, 2006). Les réformes de modernisation administrative suggèrent des implications majeures en matière d’imputabilité, notamment le changement des structures et l’attention croissante aux critères de performance (Jacob, 2009; Romzek, 2000). Dans cette veine, en épousant les préceptes de la nouvelle gestion publique (Aucoin, 1990; Hood, 1991, 1995; Hood et Peters, 2004; Jacob, 2006; Varone et Jacob, 2004), les gestionnaires publics s’intéressent...
à la mesure de la performance dans le but de contrôler, d'évaluer et de motiver le personnel; d'analyser leur budget; de promouvoir l'organisation; de célébrer les succès; d'apprendre et d'accroître la performance organisationnelle (Cantelli, Jacob, Genard et de Visscher, 2006; Behn, 2003; Charest, 2012; Hughes, 1992). De même, les agents publics de première ligne perçoivent la performance à travers des éléments de bonne gouvernance relativement à l'efficacité et au rendement fonctionnel (Emery, Wyser, Martin et Sanchez, 2008). La recherche de la performance d'une organisation engendre bien souvent des réflexions et des questionnements sur l'évaluation. Parmi les questions qui se posent fréquemment, citons « L'évaluation est-elle possible pour toutes les organisations publiques? », « Les spécificités de notre mission sont-elles compatibles avec les exigences d'évaluation et de mesure de la performance? ». Pour répondre à ces questions, nous prenons comme objet d'étude l'Ombudsman parlementaire1 qui est « un service public indépendant et impartial avec l'autorité et la responsabilité de recevoir, d'investiguer ou de résoudre de façon formelle ou informelle les plaintes contre les actions gouvernementales. Et, lorsque c'est le cas, de formuler des conclusions et des recommandations » (United States Ombudsman Association, 2003, p. 1, trad.).

Le choix de cette institution s'explique par le fait qu'il s'agit d'une organisation qui ne produit pas des biens et services publics « classiques » comme le font les ministères et organismes. L'évaluation de l'Ombudsman nécessite, par conséquent, d'entreprendre une réflexion sur les enjeux et les critères particuliers à considérer pour apprécier les interventions d'un organisme de protection des droits. Évaluer des organisations de ce type représente bien souvent des défis, car les résultats produits ne sont pas nécessairement observables en utilisant les critères couramment employés pour apprécier la performance organisationnelle. Cette difficulté peut même conduire à des erreurs dans la compréhension des activités évaluées, dans la mesure de ces activités et par conséquent dans les conclusions de l'évaluation (Eliadis, 2011).

Afin de répondre à notre question de recherche sur l'évaluabilité des activités d'un Ombudsman parlementaire, nous utilisons la typologie des organisations de Wilson (1989) qui postule que la mesure de la performance d'une organisation dépend de la nature « observable ou non » de ses résultats et de ses effets. Wilson distingue ainsi : (a) les organisations de production (résultats et effets observables), telles que les services postaux et la collecte de l'impôt; (b) les organisations professionnelles (effets observables, mais résultats non observables), telles que les hôpitaux et les armées en période de guerre; (c) les organisations procédurales (effets non observables, mais résultats observables) à l'image des armées en période de paix et les pénitencières; (d) les organisations adaptables (effets et résultats non observables) comme la diplomatie et la recherche.

Au regard de sa logique d'intervention, l'Ombudsman peut être qualifié d'organisation procédurale. D'un côté, ses résultats, notamment la correction des griefs individuels et/ou collectifs, sont observables. De l'autre côté, les effets de l'activité de l'Ombudsman ne sont pas facilement observables. En réalité, l'Ombudsman traite de la justice et de l'injustice, qui sont des notions intangibles.

Cet article offre une synthèse des connaissances sur la mesure de la performance et l'évaluation de l'Ombudsman. À notre connaissance, c'est la première fois qu'une telle démarche est entreprise. À la suite des travaux de plusieurs auteurs (Danet, 1978; Ayeni, 1993; Male, 1999; Stuhmcke, 2006; Fowlie, 2005, 2008), nous ouvrons la voie à de nouvelles réflexions théoriques et pratiques.


MÉTHODOLOGIE

Notre recherche synthétise les connaissances relatives à l'évaluation de la performance de l'Ombudsman et identifie les enjeux et les défis à prendre en considération lorsque l'on s'interroge sur cette performance.


L’ombudsman face aux défis de l’évaluation


Suite à cette recherche documentaire, nous avons répertorié 114 documents. Après avoir éliminé les doublons, nous avons effectué un premier tri sur les 72 publications restantes. Pour ce tri, nous avons lu le résumé et parcouru l’article pour vérifier l’adéquation de la publication avec les objectifs de notre étude, c’est-à-dire une publication qui a pour objet d’étude l’Ombudsman et qui contient des réflexions théoriques ou empiriques de l’évaluation de la performance de cette institution. Cette étape nous a permis d’identifier 10 documents pertinents. Nous

Figure 1 : Démarche de recherche bibliographique

doi: 10.3138/cjpe.31158
avons parcouru les bibliographies de ces 10 documents afin de repérer d'autres publications qui n’avaient pas été identifiées dans les banques de données. Cette étape nous a permis d’ajouter 15 publications à notre échantillon initial. Ainsi, notre corpus d’analyse regroupe 25 publications.

Nous avons extrait de notre corpus les informations pertinentes à notre étude. Il s’agit des connaissances qui nous informent sur la définition des concepts, les rôles et missions des Ombudsmans, la raison d’être, l’objet, les étapes et les modèles de l’évaluation, les parties prenantes concernées par l’évaluation et leur centre d’intérêt, les critères/dimensions, les variables et les indicateurs de mesure, les sources de données et les techniques d’analyse, les enseignements à tirer des évaluations, les obstacles et les éléments favorables à l’évaluation et enfin, quelques résultats observés dans la pratique de l’évaluation des Ombudsmans (Jacob, Montigny et Affodégon, 2014). Le tableau 1 présente les caractéristiques de notre échantillon de littérature.

**Tableau 1 : Caractéristiques des publications retenues dans notre échantillon (n=25)**

<table>
<thead>
<tr>
<th>Auteurs</th>
<th>Années</th>
<th>Titre</th>
<th>Type de document</th>
<th>Type d’Ombudsman</th>
<th>Théories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aufrecht, S. et M., Hertogh</td>
<td>2000</td>
<td>Evaluating Ombudsman systems In Righting Wrongs…</td>
<td>COC</td>
<td>Classique</td>
<td>Réflexion théorique</td>
</tr>
<tr>
<td>Ayeni Victor</td>
<td>1993</td>
<td>Evaluating Ombudsman Programmers</td>
<td>COC</td>
<td>Classique</td>
<td>Réflexion théorique 1</td>
</tr>
<tr>
<td>Danet Brenda</td>
<td>1978</td>
<td>Toward a Method to Evaluate the Ombudsman Role</td>
<td>A</td>
<td>Classique</td>
<td>Vérification empirique 2</td>
</tr>
<tr>
<td>Estes, C. L., 2010 S. P. Lohrer et coll.</td>
<td>2010</td>
<td>Factors Associated With Perceived Effectiveness of Local Long-Term Care Ombudsman Programs</td>
<td>A</td>
<td>Exécutif</td>
<td>Analyse quantitative 3</td>
</tr>
<tr>
<td>Fortier Isabelle Frank Fowlie</td>
<td>2009</td>
<td>Quebec Ombudsman’s Evolution…</td>
<td>COC</td>
<td>Classique</td>
<td>Statistique descriptive 3</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>A Blueprint for the Evaluation of an Ombudsman’s Office…</td>
<td>T</td>
<td>Classique</td>
<td>Revue de littérature, Approche mixte de recherche 3</td>
</tr>
<tr>
<td>Auteurs</td>
<td>Années</td>
<td>Titre</td>
<td>Type de document</td>
<td>Type d'Ombudsman</td>
<td>Méthodes</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Gilligan E.</td>
<td>2010</td>
<td>The Human Rights Ombudsman in Russia…</td>
<td>A</td>
<td>Classique</td>
<td>Approche mixte de recherche</td>
</tr>
<tr>
<td>Hannigan, A. John</td>
<td>1977</td>
<td>Newspaper Ombudsman and Consumer Complaints…</td>
<td>A</td>
<td>Exécutif</td>
<td>Analyse quantitative</td>
</tr>
<tr>
<td>Hollister, B. et C. L.</td>
<td>2013</td>
<td>Local Long-Term Care Ombudsman Program Effectiveness…</td>
<td>A</td>
<td>Exécutif</td>
<td>Analyse quantitative</td>
</tr>
<tr>
<td>ICANN Office of the Ombudsman</td>
<td>2005</td>
<td>Results Based Management Accountability Framework</td>
<td>R</td>
<td>Exécutif</td>
<td>Approche mixte de recherche</td>
</tr>
<tr>
<td>Levine-Finley, S. et Carter, J. S.</td>
<td>2010</td>
<td>Then and now: interviews with expert U.S. organizational ombudsmen.</td>
<td>A</td>
<td>Non spécifié</td>
<td>Approche mixte de recherche</td>
</tr>
<tr>
<td>Male Barbara</td>
<td>1999</td>
<td>Assessing Ombudsman Performance: Two Case Studies in North America</td>
<td>T</td>
<td>Classique</td>
<td>Revue de littérature</td>
</tr>
<tr>
<td>Male Barbara</td>
<td>2000</td>
<td>Assessing Ombudsman performance</td>
<td>COC</td>
<td>Exécutif</td>
<td>Revue de littérature</td>
</tr>
<tr>
<td>Manninen, Timo</td>
<td>2005</td>
<td>Independent Assessment of the Ombudsman Institution…</td>
<td>R</td>
<td>Classique</td>
<td>Analyse qualitative</td>
</tr>
</tbody>
</table>

(Continued)
Tableau 1 : Continued

<table>
<thead>
<tr>
<th>Auteurs</th>
<th>Années</th>
<th>Titre</th>
<th>Type de document</th>
<th>Type d’Ombudsman</th>
<th>Méthodes Théories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marin André et Gareth Jones</td>
<td>2011</td>
<td>Measuring Ombudsman Performance…</td>
<td>R</td>
<td>Classique</td>
<td>Réflexion théorique</td>
</tr>
<tr>
<td>Mohammmed F. Mahmoud</td>
<td>1991</td>
<td>Scrutiny and accountability of the public sector in middle eastern countries…</td>
<td>A</td>
<td>Classique</td>
<td>Analyse quantitative</td>
</tr>
<tr>
<td>Stuhmcke Anita</td>
<td>2006</td>
<td>Evaluating Ombudsman…</td>
<td>COC</td>
<td>Classique</td>
<td>Analyse quantitative, 5 et 6</td>
</tr>
<tr>
<td>Stuhmcke Anita</td>
<td>2012</td>
<td>Evaluating the effectiveness of an Ombudsman…</td>
<td>C</td>
<td>Classique</td>
<td>Réflexion théorique</td>
</tr>
<tr>
<td>Tyler, Harrison</td>
<td>2004</td>
<td>What is success in the Ombudsman Process?</td>
<td>A</td>
<td>Classique</td>
<td>Approche mixte de recherche, 2</td>
</tr>
<tr>
<td>Van Roosbroek, S. et S. Van de Walle</td>
<td>2008</td>
<td>The Relationship between Ombudsman, Government…</td>
<td>A</td>
<td>Classique</td>
<td>Approche mixte de recherche, 6</td>
</tr>
<tr>
<td>Olesen, G. et Sakashvili, N.</td>
<td>2006</td>
<td>Defending Human Rights in Georgia…</td>
<td>R</td>
<td>Classique</td>
<td>Approche mixte de recherche</td>
</tr>
<tr>
<td>United States Ombudsman Associ-ation</td>
<td>2003</td>
<td>USOA Governmental Ombudsman Standards.</td>
<td>G</td>
<td>Classique</td>
<td>-</td>
</tr>
</tbody>
</table>

Légende :
1 = Gestion axée sur les objectifs, 2= Sociologie des plaintes, 3 = Gestion axée sur les résultats, 4 = Choix rationnel, 5 = Mesure axée sur la performance, 6= Mesure axée sur l’impact, 7 = Mesure axée sur les normes de qualité.
Raison d’être de l’Ombudsman : les finalités de son évaluation


Sur la base de l’analyse de notre échantillon, il ressort que la principale mission de l’Ombudsman parlementaire est de combattre « l’injustice administrative ». Celle-ci se manifeste dans des actions non conformes à la loi ou des interventions publiques oppressives ou discriminatoires (Mohammed, 1991; Thomas, Martin et Kirkham, 2013). Face à ces injustices, un citoyen peut formuler une plainte et demander réparation. La formulation d’une plainte résulte d’un processus en trois étapes. Une expérience préjudiciable non perceptible doit être transformée en situation préjudiciable perceptible (naming). Cette situation perceptible doit être transformée en grief (blaming). Dans la manifestation du grief, l’organisation est...

L’Ombudsman néglige pas à l’impératif d’imputabilité qui s’est généralisé dans la fonction publique et qui se matérialise à travers les exercices d’évaluation et de reddition de comptes (Bovens, Goodin et Schillemans, 2014). À ce sujet, Stuhmcke (2006) considère que l’évaluation s’inscrit dans le développement naturel de cette institution lorsqu’il affirme qu’« après la période initiale de prescription, c’est-à-dire la période durant laquelle tout le monde voulait s’assurer de la nécessité d’avoir un Ombudsman, suivie de la période de description au cours de laquelle on a délimité ses fonctions et juridictions, vient maintenant le temps d’un engagement sérieux envers l’évaluation » (p. 25, trad.).

Cette évaluation poursuit des ambitions techniques et politiques (Aufrecht et Hertogh, 2000). L’évaluation est technique puisqu’elle concorde avec les exigences de la gestion publique contemporaine qui oblige chaque gestionnaire ou chaque titulaire de charge publique à rendre des comptes de leur gestion (Aucoin et Heintzman, 2000; Jacob, 2009). C’est ainsi que l’Ombudsman parlementaire présente un bilan annuel de gestion qui permet de documenter les décisions du titulaire relatives à l’allocation des ressources (humaines, financières, matérielles et informationnelles). Dans cette veine, l’évaluation est perçue comme un outil de pilotage, c’est-à-dire un processus continu qui s’appuie sur des données collectées au fur et à mesure par un dispositif interne de suivi de la performance de l’Ombudsman. L’évaluation technique est principalement centrée sur les critères de pertinence et d’efficacité appliqués aux processus et aux pratiques de gestion interne. En améliorant sa gestion, l’Ombudsman veille à satisfaire les besoins de sa « clientèle » qui peut indirectement rejaillir sur la satisfaction à l’égard du travail de l’Ombudsman en particulier et de l’institution en général. Le critère de la satisfaction est par la suite fréquemment repris dans les modèles d’évaluation de l’Ombudsman, même s’il se concentre davantage sur les processus de résolution des plaintes que sur les résultats ou les effets du travail de l’Ombudsman. Par conséquent, l’évaluation technique met peu l’accent sur les décisions révisées ou les modifications structurelles qu’apportent les organisations publiques suite à une intervention de l’Ombudsman.

servent ensuite d’argument dans l’arène politique. Cette évaluation insiste plutôt sur des critères d’intégrité et de compétences que sur des critères économiques ou managériaux (Aufrecht et Hertogh, 2000; Fowlie, 2008). De plus, cette évaluation, centrée sur le critère de conformité, ne prend pas en considération la mission première de l’Ombudsman qui est de combattre l’injustice administrative.

**Vers une congruence des modèles d’évaluation de l’Ombudsman**

La littérature qui s’intéresse à l’évaluation de l’Ombudsman identifie cinq grands modèles d’évaluation que nous présentons brièvement ci-dessous avant de nous intéresser aux défis et aux enjeux relatifs à l’évaluation de la performance de l’Ombudsman.

Primo, le modèle théorique d’évaluation axé sur la sociologie des plaintes évalue les dimensions clientèle, administration publique et bureau de l’Ombudsman (Danet, 1978). Les évaluations se concentrent principalement sur les activités et les résultats concernant la résolution ou non des plaintes (voir figure 2). L’une des caractéristiques de ce modèle est qu’il s’appuie sur l’évolution du volume de plaintes traitées pour évaluer la dimension « clientèle ». Cependant, un volume important de plaintes ou une diminution du nombre de plaintes enregistrées ne sont pas des mesures suffisantes pour apprécier pleinement la performance de cette organisation (Fortier, 2009; Hannigan, 1977). En effet, une augmentation du volume de plaintes d’une année à l’autre peut reflêter une augmentation des injustices administratives ou une amélioration de la notoriété ou de l’accessibilité à l’Ombudsman. Par exemple, lorsqu’il a été créé en 1998, le médiateur russe a reçu 7 000 plaintes. Il en a reçu 22 000 en 1999 et 56 729 en 2007. Au cours de cette période, le taux de résolution des plaintes demeure constant et se situe aux environs de 15 à 20 % (Gilligan, 2010). De la même manière, une baisse des plaintes peut signifier un désintérêt ou une méconnaissance des citoyens par rapport au travail de l’Ombudsman ou signifier tout simplement une amélioration dans la prestation des services publics. Dans ce sens, Hannigan (1977) rapporte qu’environ 91 % des plaignants tentent de négocier avec l’agence en cause avant de se plaindre à l’Ombudsman. De même, environ 28 % de plaignants ont eu recours d’abord à une tierce personne avant de solliciter l’Ombudsman. Cette forme d’évaluation exclut donc le volet prévention et le changement systémique résultant des interventions de l’Ombudsman.

Sur un autre plan, le modèle suggère l’adoption des recommandations de l’Ombudsman par les administrations publiques comme un indicateur reflétant la performance. Cependant, les spécialistes de l’administration publique savent que l’adoption et la mise en œuvre d’une recommandation peuvent entraîner des réformes substantives ou procédurales qui dépendent aussi en grande partie de la volonté politique des gouvernants et des intérêts des administrateurs concernés. La mise en œuvre des recommandations n’est donc pas le résultat de la seule intervention de l’Ombudsman. Ce modèle documente aussi la performance du bureau de l’Ombudsman à travers la mesure de l’efficience du personnel sur la base du nombre de dossiers traités par les employés. L’efficience du bureau est reflétée par
Figure 2: Synthèses des dimensions et composantes de l’évaluation
la vitesse de traitement des dossiers ainsi que le nombre de cas non résolus qui subsistent à la fin d’un exercice. Ce critère ne tient pas compte de la complexité des dossiers et s’éloigne de l’acceptation répandue de l’efficience qui met en relation les ressources utilisées et les effets produits par un programme ou une politique. Bien qu’apportant un apport inestimable à la compréhension des multiples aspects évaluatifs de l’Ombudsman, cette approche d’évaluation manque de précision sur les indicateurs d’effet et d’impact de l’Ombudsman. Elle occulte aussi les dimensions administratives et politiques de l’Ombudsman.


Discussion : des défis et des enjeux à l’évaluation de la performance de l’Ombudsman parlementaire

Les cinq modèles théoriques et empiriques que nous venons de présenter évaluent les intrants, les processus et les résultats des activités et des interventions
de l’Ombudsman. Ils omettent en général, la mesure des effets et l’impact direct de l’institution sur l’administration publique, les citoyens et les parlementaires comme nous le voyons dans la figure 2.

En observant ces dimensions et composantes de l’évaluation, nous constatons que la définition de la portée de l’évaluation de l’Ombudsman représente un premier défi. La figure 2 présente les critères de pertinence, d’efficacité, d’efficience, de coût-efficacité, d’utilité et de durabilité pour évaluer l’Ombudsman parlementaire. En considérant que l’Ombudsman ne génère a priori pas de revenus directs (Stuhmcke, 2012), il est difficile de pouvoir entreprendre une analyse économique ou celle-ci conclurait vraisemblablement que ce service n’est pas rentable. En effet, l’Ombudsman n’est pas une organisation de production de biens et services dont les réalisations ou les effets sont monétarisables. Les données financières donneront à l’évaluateur un aperçu sur la gestion financière de l’Ombudsman, mais seront de peu d’utilité pour apprécier l’économie ou l’efficience de l’organisation (Ayeni, 1993; Fowlie, 2007, 2008; United States Ombudsman Association, 2003).

L’Ombudsman parlementaire opère dans un sous-système composé d’acteurs politiques et administratifs ainsi que de plaignants en particulier et de citoyens en général. Les comportements et attitudes de ces acteurs ont une influence directe sur les activités et le fonctionnement de l’Ombudsman (Ayeni, 1993). Par exemple, nous savons que les citoyens qui se plaignent ont généralement un niveau d’éducation plus élevé et occupent une fonction plus prestigieuse que la moyenne des citoyens (Hannigan, 1977; Van Roosbroek et Van de Walle, 2008). De plus, nous sommes face à une organisation qui travaille très souvent de façon informelle en recourant à une approche très flexible. Cette organisation du travail de l’Ombudsman et la relative difficulté à entreprendre la traçabilité de ses interventions représentent des défis lorsque vient le temps d’apprécier sa performance (Stuhmcke, 2012).

l’évaluateur accorde peu de place à l’innovation et les données disponibles sont essentiellement centrées sur les produits (Van Dooren, Bouckaert et Halligan, 2010). La deuxième méthode consiste à collecter des données primaires à l’aide de sondages ou d’entrevues auprès des usagers des services de l’Ombudsman ou de personnes ressources (Ayeni, 1993; Fowlie, 2008; Stuhmcke, 2006; Olesen et Sakashvili, 2006). Les administrateurs réceptifs et les plaignants constituent une source primaire d’informations. La nature et la qualité des données dépendent des aptitudes de l’évaluateur à surmonter les biais méthodologiques. Par exemple, un échantillon constitué uniquement de plaignants satisfaits de la résolution de leur plainte affecterait l’évaluation de la qualité et de l’efficacité des services. Ainsi, une méthode mixte d’évaluation (utilisation de données qualitatives et quantitatives provenant de rapports d’activités et de sondages et/ou d’interviews) se prête mieux à l’évaluation de l’Ombudsman parlementaire (House, 1994; Reichardt et Rallis, 1994; Rossi, 1994; Estes et coll., 2010; Hollister et Estes, 2013). À titre d’illustration, nous avons vu que l’approche d’évaluation axée sur la sociologie des plaintes se prêterait bien à une évaluation recourant à des méthodes quantitatives (nombre de plaintes reçues, délais de traitement, taux de résolution, etc.). L’analyse de ces seules données quantitatives ne permettrait pas à l’évaluateur d’en faire une interprétation qui rendrait compte avec précision des facteurs contextuels qui peuvent influencer les activités de l’Ombudsman (ex. : contexte politique propice à des abus de pouvoir, degré de complexité des plaintes, etc.).

Par ailleurs, indépendamment de la méthode de collecte des données, l’évaluateur tiendra compte de la confidentialité qui caractérise les activités de l’Ombudsman. La confidentialité étant un standard de service des Ombudsmans qui prescrit la protection de la source de l’information. L’Ombudsman ne peut révéler une information lorsque la confidentialité est garantie par le processus de médiation ou lorsque la loi l’interdit. Il est évident que la qualité et la validité de l’évaluation seront affectées si l’évaluateur ne peut avoir accès aux données relatives aux plaignants. Dans l’étude de cas de Tyler (2004), des plaignants ont refusé de participer à une enquête. Les raisons invoquées concernaient la nature des sujets, le lieu des interviews et le mode de sélection des interviewés qui ne garantissaient pas la confidentialité.

Un autre défi de l’évaluation est l’intégration des attentes et des préoccupations de nombreux acteurs concernés par l’évaluation de l’Ombudsman. Les modèles d’évaluation présentés accordent une prépondérance à quelques acteurs, tels que l’Ombudsman, les autorités gouvernementales et législatives, les administrations publiques et les plaignants. L’Ombudsman est directement concerné par une démarche évaluative qui lui permet de mesurer sa performance afin d’améliorer ses services. De leur côté, les autorités gouvernementales et législatives constituent également des acteurs clefs, car, non seulement, elles s’intéressent à sa gestion, mais elles souhaitent aussi connaître ses impacts sur l’administration publique en particulier et sur la société en général (Aufrecht et Hertogh, 2000; Danet, 1978; Fortier, 2009; Mohammed, 1991; Stuhmcke, 2012). Les administrations publiques contre lesquelles les plaintes des usagers sont dirigées s’intéressent elles
aussi à l'évaluation de l'Ombudsman. Dans ce cas, il est possible de distinguer les « administrateurs réceptifs » qui perçoivent l'Ombudsman comme un acteur qui les aide à améliorer leur service (Aufrecht et Hertogh, 2000) des « administrateurs récalcitrants » qui perçoivent l'Ombudsman comme une menace à l'exécution de leurs missions ou une nuisance à leur réputation. Il arrive que ces derniers soient tentés d'instrumentaliser leur participation à l'évaluation pour régler des comptes avec l'Ombudsman et tenter d'entacher sa crédibilité (Aufrecht et Hertogh, 2000). Enfin, les plaignants sont des acteurs incontournables dans une évaluation de l'Ombudsman puisqu'ils peuvent être une source d'information précieuse pour un évaluateur et, dans certains cas, ils peuvent également avoir observé des changements de pratiques administratives suite à la résolution de leur plainte (Aufrecht et Hertogh, 2000; Danet, 1978; Mohammed, 1991; Stuhmcke, 2012). Dans les faits, le processus évaluatif est bien souvent plus ouvert et rejette également les préoccupations d'autres acteurs. Il s'agit par exemple du public en général, des médias, des groupes d'intérêts, des institutions internationales, d'autres ombudsmans ainsi que les opportunistes, qui « gagnent leur vie » grâce au mauvais fonctionnement de l'administration publique et les concurrents à l'Ombudsman, c’est-à-dire les autres services de résolution de conflits, tels que les groupes de défense et les institutions d’audit (Behn, 2003; Van Dooren, Bouckaert et Halligan, 2010).

CONCLUSION

Suite à cette revue de la littérature, nous constatons que l'évaluation de la performance de l'Ombudsman est un processus complexe en raison de la nature procédurale des missions confiées à cette organisation. De notre analyse, il ressort que l'appréciation de la performance décrite dans les modèles d'évaluation recensés met l'accent sur la mission d'arbitrage et de gestion des plaintes et délaisSENT d'autres missions de l'Ombudsman, telles que la promotion d'une bonne administration ou l'éducation citoyenne. L'appréciation de la performance de l'Ombudsman est parfois assimilée à la Quête du Graal (Marin et Jones, 2011). Cette expression reflète bien les difficultés que peuvent éprouver les évaluateurs lorsqu'ils cherchent à apprécier l'intangible. Toutefois, il ne faudrait pas en déduire qu'il est impossible d'évaluer la performance de l'Ombudsman y compris dans les dimensions « intangibles » de sa mission. À cet égard le modèle d'évaluation d'impact offre le plus de potentiel pour déceler les changements systémiques engendrés par une intervention de l'Ombudsman ou la contribution de celui-ci à la qualité globale des services publics offerts à la population. En fait, l’évaluation de l’Ombudsman nous fait penser à l’histoire de l’individu qui cherche ses clés sous un lampadaire, non pas parce qu’il les a perdues à cet endroit, mais plutôt parce que c’est le seul endroit éclairé de la rue. Sur la base d’une étude comparative des rapports annuels d’activités rédigés par 13 Ombudsmans dans 5 pays (Afrique du Sud, Australie, Canada, États-Unis, France, Nouvelle-Zélande et Royaume-Uni), il est possible d’identifier les indicateurs de performance les plus couramment utilisés (Affodegon et Montigny, 2016). Il ressort de cette analyse que les indicateurs

doi: 10.3138/cjpe.31158

Ainsi, pour retirer le plein potentiel des modèles évaluatifs que nous avons présentés, il nous semble avantageux de les aborder dans une logique de complémentarité plutôt que d’opposition. En effet, chacun traduit une partie des réalités du fonctionnement de l’Ombudsman. Par exemple, le modèle d’évaluation axée sur la sociologie des plaintes insiste sur les relations de l’Ombudsman avec les principaux acteurs qui contribuent à sa raison d’être, tels que les plaignants et l’Administration publique. L’évaluation axée sur les objectifs met l’accent sur les conditions d’existence et d’opérationnalisation des activités de l’Ombudsman. Le modèle d’évaluation axée sur les résultats apprécie principalement la gestion interne, c’est-à-dire la capacité à produire des résultats conformément au plan de développement stratégique et à la déclaration de service aux citoyens. L’évaluation d’impact direct renvoie à la satisfaction directe de la clientèle. Le modèle d’évaluation axée sur les normes de qualité est une perspective assez intéressante pour les Ombudsmans parlementaires. Cette évaluation de l’Ombudsman gagnerait aussi à être inclusive pour tenir compte de la diversité des besoins et des intérêts des parties prenantes.

La description que nous faisons de l’évaluation de l’Ombudsman s’appuie essentiellement sur une synthèse de connaissances de publications scientifiques. Nous sommes conscients que cela limite la portée de notre étude. Dans des recherches futures, il serait avantageux de tirer parti d’autres types de publications et, notamment, des évaluations ou rapports d’analyse de la performance qui permettraient de documenter plus finement les dimensions pratiques de l’évaluation de l’Ombudsman. La synthèse des modèles que nous présentons dans cet article pourrait alors servir de point de départ à l’analyse de ces autres sources et permettrait d’identifier quelles sont les fonctions d’un Ombudsman qui sont davantage sujettes à évaluation ainsi que les approches et méthodes appropriées pour entreprendre cet exercice. De plus, notre analyse porte principalement sur les Ombudsmans « nordiques », c’est-à-dire des organismes de résolution de conflits qui interviennent face à la mal-administration ou l’injustice administrative. En raison de l’évolution récente du mandat de certains Ombudsmans qui se consacrent spécifiquement à la promotion et à la protection des droits de l’Homme, il serait également intéressant d’étudier les répercussions que peut avoir cette nouvelle mission sur les approches, les orientations et la conduite d’une évaluation.
Remerciement : Les auteurs remercient le Protecteur du citoyen du Québec et la Chaire de recherche sur la démocratie et les institutions parlementaires de l’Université Laval pour le financement de cette recherche.

NOTE

Dans certaines juridictions, l’Ombudsman est également connu sous le nom de « Protecteur du citoyen », de « Médiateur de la République » ou encore de « Défenseur des droits ».

RÉFÉRENCES


doi: 10.3138/cjpe.31158


PRÉSENTATION DES AUTEURS

Seyive Wilfried Affodegon est doctorant en science politique à l’Université Laval (Québec-Canada). Il est titulaire d’un diplôme d’ingénieur agroéconomiste et d’un master en management. Ses intérêts de recherche incluent l’administration publique comparée, le management public, la politique publique comparée, la mesure de la performance et l’évaluation de l’action publique, le parlementarisme et les méthodes de recherche en science politique.


Éric Montigny est professeur au département de science politique de l’Université Laval, il est directeur scientifique de la Chaire de recherche sur la démocratie et les institutions parlementaires. Titulaire d’un doctorat en science politique de l’Université Laval, il dispose d’une maîtrise en administration publique de l’ENAP. Il enseigne des cours en politique canadienne et québécoise, en gouvernance ainsi que sur les partis politiques. Ses travaux de recherche traitent principalement de la démocratie interne des partis politiques, du militantisme, du parlementarisme et de la gouvernance.
Making Evaluation More Responsive to Policy Needs:
The Case of the Labour Market Development Agreements

Yves Gingras, Tony Haddad, Andy Handouyahia, Georges Awad, and Stéphanie Roberge
Employment and Social Development Canada

Abstract: This note describes how Employment and Social Development Canada evaluation staff transformed the Labour Market Development Agreement (LMDA) evaluation process to make it more timely, cost-effective, and relevant for policy development. The note provides background on the LMDAs and discusses key drivers for changing the evaluation approach. In particular, it describes the benefits of using small targeted studies, rich administrative panel data, and building in-house evaluation capacity. It concludes with some lessons learned for the evaluation practice.

Keywords: administrative data, capacity building, evaluation approach

INTRODUCTION

Departments such as Employment and Social Development Canada (ESDC) increasingly expect evaluators to produce targeted and timely information to support ongoing policy work. The need to produce more timely and relevant evaluation pushes evaluators to rethink how they do evaluation. Conducting a rigorous
evaluation usually entails several methodological, quality control, and approval steps that can make the process very lengthy. However, as reported by Bourgeois and Lahey (2014), in the recent past the lack of timeliness of evaluation in ESDC was often a barrier to its use.

The pressure to innovate also became greater with the introduction of the 2009 Policy on Evaluation (Government of Canada, Treasury Board Secretariat, 2009) that required an evaluation of all program spending every five years. On one hand, the evaluation function had to develop more capacity to meet policy requirements. On the other hand, expanding the capacity was limited by budget constraints (Bourgeois, Toews, Whynot, & Lamarche, 2013).

This particular context motivated evaluators at ESDC to rethink how they evaluated 12 federal/provincial/territorial bilateral administrative agreements. The purpose of this practice note is to describe how the bilateral approach for evaluating those 12 agreements was streamlined into one continuous evaluation process. With this new approach, evaluations on different aspects of the agreements are conducted and released simultaneously for all provinces/territories and for Canada periodically. As well, the evaluation builds on the fact that rich administrative data are available to estimate outcomes and impacts and to conduct analyses in-house.

The note first describes the LMDAs and then highlights the drivers for changing the evaluation approach. It next explains the key features of the new approach and its observed benefits. It concludes by discussing lessons learned from this experience.

**THE LABOUR MARKET DEVELOPMENT AGREEMENTS (LMDAS)**

Introduced in 1996 under Part II of the Employment Insurance Act, the LMDAs are agreements between the Government of Canada and each of the 13 provinces and territories. They include the transfer of $1.95 billion annually from the federal government for the delivery of employment programs and services that help mostly Employment Insurance-eligible unemployed individuals find and maintain employment.1

Since 2010, provinces and territories have had full responsibility over the design and delivery of LMDA-funded programs and services as long as they correspond to categories defined in the Employment Insurance Act. The categories include, for example, financial assistance for training. These categories are similar to the main type of active labour market programs delivered in other countries (Gunderson, 2003) and to those used in other federally funded employment programs.

**DRIVERS FOR RETHINKING THE EVALUATION APPROACH**

The LMDAs require a first summative evaluation in the third year of the agreement’s implementation2 and subsequent evaluations of the impact and effectiveness

doi: 10.3138/cjpe.31119

*CIPE* 32.2, 244–253 © 2017
of the programs every three to five years. Each province and territory has the choice of evaluating their LMDA on their own or jointly with the Government of Canada. Every province and territory except Quebec opted for a joint evaluation.

Twelve bilateral formative evaluations were conducted between 1998 and 2012 and were followed by 12 bilateral summative evaluations conducted between 2002 and 2012. Each summative evaluation took two to three years to complete. Because the evaluation team had the capacity to conduct two to three evaluations at the same time, it was not possible to comply with the LMDA evaluation clause and the Treasury Board’s *Policy on Evaluation* (2009). The process was lengthy and costly because each evaluation was undertaken by external contractors and relied on large surveys to collect information on program outcomes and impacts. Approximately $10 million was spent in external contracts alone.

When planning for the second cycle in 2011–2012, expanding the evaluation capacity to carry more bilateral evaluations at the same time was not an option given the budget constraints. The only realistic option was to rethink the evaluation approach to increase efficiency but also to make the evaluations more useful.

Academics and stakeholders (e.g., industry representatives) raised concerns with the accountability around the LMDAs and the fact that no national level assessment was available (Canada, 2015). Federal and provincial/territorial representatives also thought that the findings from the first round of summative evaluations were of limited use, partly because of the lack of timeliness and the limited evidence they generated about the program effectiveness in some jurisdictions. ESDC had to rely on surveys of participants and nonparticipants to estimate impacts. These surveys often resulted in small sample size that made quantitative analyses of impacts difficult. As well, to minimize recall errors, the survey followed individuals only over the short term after participation, although a fair assessment of employment program impacts requires a longer observation period. As shown in the literature, impacts of employment programs are often more positive in the medium- to long-term compared to the short-term (Card, Kluve, & Weber, 2015). Finally, it was difficult for federal officials to get an overall understanding of LMDA effectiveness as it was challenging to roll up impacts from 12 different evaluations undertaken over 10 years.

**THE NEW APPROACH**

The second round of evaluations took place between 2012 and 2016 using a new approach developed in consultation with provinces/territories and ESDC program managers and policy makers. Key features include

*Breaking the evaluation into small annual studies undertaken at the national and provincial/territorial levels simultaneously*

Two to three studies were conducted annually over a span of five years. Findings were generated for Canada overall and for each province/territory separately. One annual study measured program incremental impacts on key labour market
indicators (e.g., earnings) over the short- and medium-term after participation and for various groups of participants (e.g., youth, older workers) or time periods (i.e., short- and medium-term after participation).

One or two additional studies were conducted using targeted questions about program design, delivery, relevance, effectiveness, and efficiency, with a goal of supplementing what was already known about the LMDAs. For example, some studies examined the design and delivery of specific programs (e.g., skill training) to identify best practices and understand how they are targeted. Another study examined how the timing of participation in Employment Assistance Services affects participants’ labour market impacts. Evaluators also completed the first-ever cost-benefit analysis of LMDA programs based on six years of postprogram data. Using a similar framework as in Bloom et al. (1997), Heckman, Lalonde, and Smith (1999), and Heckman, Moon, Pinto, Savelyev, and Yavitz (2010), the analysis estimated the extent to which benefits from the program exceeded the costs six years after the end of participation.

Multilateral and bilateral governance process

A two-level governance structure was put in place to oversee the evaluation process. An Evaluation Steering Committee consisting of representatives of ESDC and the 12 participating provinces/territories was mandated to decide on the evaluation work plan as well as to approve the methodologies and the national reports. Joint Evaluation Committees including representatives of ESDC and officials from each participating province and territory were also set up to discuss issues specific to each jurisdiction.

Ongoing consultation of evaluation partners on the scope of the evaluation

Provinces/territories and federal program representatives were consulted on an ongoing basis to determine the evaluation issues to examine. Each year, members of the multilateral Evaluation Steering Committee were consulted to identify their priority subject for the upcoming studies and voted on their preferred topics.

Ongoing dissemination of evaluation findings

Evaluation findings were shared with the provinces/territories and ESDC program stakeholders as they became available. At the end of the cycle, findings from each study were summarized in a bilateral report for each jurisdiction.

Relying on administrative data as the main source of information for the evaluation

The evaluation team capitalized on the fact that good quality administrative data about program beneficiaries were available at ESDC. The team created a longitudinal database for evaluation purposes. It included data from the Employment Insurance claims, program participation, and taxation files from the Canada Revenue Agency over a period of 20 years. It was updated annually with the most recent data available. It covered 100% of LMDA participants and up to 20% of EI claimants who never participated in the LMDAs. The latter data were used to
create comparison groups for nonexperimental analyses of program incremental impacts. Overall, this database provided a wealth of historical information on participants’ and nonparticipants’ earnings, labour market attachment, and sociodemographic characteristics before and after participation.

Relying on administrative data to carry out the LMDA evaluation was possible because taxation data became accessible to ESDC evaluators toward the end of the first LMDA evaluation cycle. As well, over the years, the department has adopted strict rules and processes to maintain the privacy and security of the data. For example, data used for evaluation do not contain personal information such as Social Insurance Numbers, names, and addresses. Those are either removed from the file or masked with sequence numbers.

Under the new approach, administrative data were used to estimate program incremental impacts over up to five years after participation. Evaluators used propensity score matching techniques as in Blundell and Dias (2002), Gerfin and Lechner (2002), and Sianesi (2004) and estimated program impacts using difference-in-differences as in Heckman, Ichimura, and Todd (1997) and Heckman et al. (1999). For validation purposes, results were produced with three matching techniques (Kernel Matching, Inverse Probability Weighting, and Nearest Neighbours), and the models were tested with sensitivity analyses (e.g., \texttt{rbound} and \texttt{mhbounds} Stata command provided by DiPrete & Gangl, 2004, and Becker & Caliendo, 2007). Academic experts were hired to review the methodology and findings. Using propensity score matching techniques is common in evaluations of employment programs. They represent accessible alternatives to randomized control trials and were found to produce generally similar results as those from randomized control trials (Card, Kluve, & Weber, 2015).

Administrative data were also used to produce various statistics on the sociodemographic profile of participants and their labour market outcomes over time. However, administrative data provide limited information on how the programs are designed, delivered, and targeted. Qualitative methods (e.g., key informant interviews) were used to examine those issues and explain the quantitative results.

**Conducting the Evaluations In-house**

All LMDA evaluation studies were conducted in-house by two teams. One team of three methodologists specialized in data processing and econometric analyses. This team was responsible for all questions related to data, methodologies, and quantitative analyses. Another team of three evaluators was responsible for the qualitative fieldwork, coordinating evaluation committees, maintaining relationships with partners, writing evaluation reports, and communicating findings.

**BENEFITS FROM THE NEW APPROACH**

Several benefits emerged from the new evaluation approach, including

*Better supporting the policy development process.* The scope of the evaluation was continuously adapted to better serve the needs of program managers and
policy makers. Selecting the study subjects annually provided an opportunity to address the emerging evidence needs of evaluation users and to fill knowledge gaps identified in previous studies. Evaluation results were shared on an ongoing basis with key evaluation users through technical reports as well as regular presentations to the Performance Measurement and Evaluation Committee and to provinces/territories. Because of this flexibility and the ongoing sharing of information, evaluation results were used to support policy work on the future on the labour market transfers. As well, findings from the study on the impacts of early participation in Employment Assistance Services contributed to inform a new departmental initiative on early targeting of programs to unemployed individuals.

Regrouping efforts and sharing best practices. Working at the multilateral level gave an opportunity to provinces and territories to regroup efforts and exchange on best practices. In the first evaluations, quantitative analyses were often limited by the small number of participants in some of the smaller provinces/territories. In the second cycle, when particular analyses were not possible, the province or territory was able attain insights from the national or regional level analyses (e.g., Atlantic). The national studies also highlighted best practices across Canada in terms of program design and delivery.

Producing highly credible and robust evidence on program incremental impacts. The quality and wealth of information in the administrative data coupled with the evaluation team capacity to apply state-of-the art econometric techniques generated robust estimates of program effects. With that regard, in 2014 the evaluation team invited ESDC officials to attend a panel of academic experts mandated to review methodologies and results. This panel confirmed the reliability of the impact estimates and helped evaluation users gain confidence in the value of the evaluation.

Demonstrating the value of collecting good data. At ESDC, evaluations of most labour market programs (e.g., programs targeted to youth or Employment Insurance benefits) rely on administrative data to examine impacts and/or outcomes. In fact, any programs that collect data on beneficiaries can potentially be evaluated using the same methods as those used for the LMDA evaluation. The LMDA evaluation provides another convincing example that investing in good quality data can lead to important payback. The use of administrative data produces robust evidence on program performance in a timely and efficient manner.

Reducing the cost of undertaking evaluation. Undertaking the work in-house helped reducing evaluation costs. External contracts were limited to peer reviews and advice on methodology and data processing. These expenses approximated $70,000 per year compared to about $1 million per year for the first cycle.

CONCLUSION

At the time of writing this note, the second evaluation cycle was winding down and evaluators had had the opportunity to reflect on the lessons learned for the evaluation practice.
Moving from a model where the evaluations are contracted to one where evaluations are undertaken in-house necessitated the development of internal capacity. It was crucial to build teams who had the technical skills to keep up with recent methodological developments and to implement innovative techniques. Building capacity required time and resources for research and training on methodologies.

Dividing the evaluation work across two specialized teams had some advantages since it allowed each team to develop expertise in its own area of work. The methodologists could focus on improving data processing and implementing innovative techniques, while the more generalist team could focus on ways to better communicate evaluation results. This division of work also had its challenges. It took some time for both teams to find an efficient way to coordinate and to communicate effectively about the expectations and constraints they are facing. For example, it is not because a certain type of analysis is feasible from a methodological standpoint that it is necessarily relevant from a policy standpoint.

Investing in the right tools, software, and computational capacity was also important. Having access to administrative data can allow quick analyses, but this work can only be done as quickly as technology allows, underscoring the vital importance of accessing appropriate technology.

As proposed by Michael Quinn Patton, evaluation that focuses on utilization should be designed for and with intended users (Patton, 2008). Directly involving all intended users of LMDA evaluation findings would have been nearly impossible, since LMDA programs are sometimes delivered by more than one department in the provinces/territories. Still, the success of this project in making evaluation more relevant to policy was largely due to having a governance process with clear roles and responsibilities and the ongoing consultation of evaluation committee members to ensure that all the players had an opportunity to express their interests. It was necessary to build in flexibility to allow committee members to conduct internal consultations on evaluation deliverables.

Toward the end of the cycle, provinces and territories as well as federal program representatives were surveyed to collect their views on the evaluation approach. They generally expressed a high level of satisfaction with the approach and its outcomes. The main concern noted was that the needs and interests of the small provinces/territories tend to get lost in the multilateral process.

Finally, having access to quality administrative data was critical, as it allowed the team to measure program incremental impacts and to do so in a timely and cost-effective manner. The LMDA evaluation was privileged in the sense that data on program participation and the labour market history before and after participation were available.

While this evaluation model cannot necessarily be replicated to other programs as is, some of its key features may be a source of inspiration for the evaluation of other federal/provincial/territorial agreements or horizontal initiatives. The third cycle of LMDA evaluation will build on the successes and lessons learned in this second cycle approach.
ACKNOWLEDGEMENTS

In developing methodologies and analyses for LMDA evaluations, the evaluators benefited from advice and peer reviews from various academic experts. As well, in 2014, the evaluators held an internal expert panel to critically review and discuss the methodology for the incremental impact analysis. In particular, we would like to thank Professors Walter Nicholson, Jeff Smith, Guy Lacroix, and David Gray for providing advice on the LMDA evaluation studies.

NOTES

1 LMDA programs are offered only to Employment Insurance-eligible individuals with the exception of Employment Assistance Services, which are also provided to non-Employment Insurance eligible individuals.

2 Agreements are available online at https://www.canada.ca/en/employment-social-development/programs/training-agreements/lmda.html

3 The agreements include data exchange provisions for the transfer of provincial/territorial data on the participants and the programs in which they participated (see Labour Market Development Agreements, retrieved from https://www.canada.ca/en/employment-social-development/programs/training-agreements/lmda.html).

REFERENCES


**AUTHOR INFORMATION**

**Yves Gingras** has been the Director General of the Evaluation Directorate at Employment and Social Development Canada since 2013. He studied economics at Laval University, and earned his MA in 1992 with a fellowship from Natural Resources Canada. Prior to this, he studied at the Royal Military College of Saint-Jean. In 1992 he joined the Public Service of Canada through the Accelerated Training Program for Economists. Subsequently, he worked as a researcher at HRDC and as a policy advisor at the Priorities and Planning Secretariat at the Privy Council Office. In 2004 he joined Finance Canada as a chief in the Personal Income Tax Division. He was appointed Senior Director in 2009 and interim DG of the Labour Market Policy Directorate at HRDC in 2010. He represented Canada at the Employment, Labour and Social Affairs Committee of the OECD. He also did research at Laval University, the Central Bank of Finland, and the Central Bank of Malta. He has published articles on energy and labour market issues.

**Tony Haddad** completed his PhD in sociology at York University in 1998. He was the Director of the Partnership Evaluation Division from 2006 to 2014, during which he led the evaluation work on labour market programs funded by ESDC and delivered by provinces,
territories, and organizations, including Indigenous communities. Under his lead, the Partnership Evaluation Division developed and refined the use of linked program administrative data to undertake evaluation work that was more timely, less resources intensive, and very statistically robust. He is currently leading the Indigenous Labour Market Information team within the Indigenous Affairs Directorate, Skills and Employment Branch, in ESDC.

Andy Handouyahia has been working in data science for more than 25 years. He is currently a manager in the Evaluation Directorate at Employment and Social Development Canada where he is responsible for data and methodology. He earned an MA in science in 1997 from Université de Sherbrooke. He has been giving lectures at Université du Québec en Outaouais for approximately 20 years. He has been responsible for data collection and management at Statistics Canada for more than 13 years and has also worked at the Treasury Board Secretariat of Canada as a manager of business solutions.

Georges Awad has an MA in economics from the Université de Montréal. After working two years as an economist in Emploi-Québec, he joined the Evaluation Directorate of Employment and Social Development Canada in April 2001. He continues to occupy the position of an evaluation manager since 2005, working on the evaluation of various labour market programs.

Stéphanie Roberge is a senior program evaluator. She started working in evaluation at Canadian Heritage in 2006. She joined Employment and Social Development Canada few months later and has been working in the Evaluation Directorate since then. She earned an MA in political science with a specialization in public policy and public administration from Université de Montréal in 2005.
Using Rubrics for an Evaluation:
A National Research Council Pilot

Ghislaine Tremblay and Melissa Fraser
Office of Audit & Evaluation, National Research Council Canada

Frédéric Bertrand
Independent Senior Evaluation Consultant, Ber /

Abstract: Rubrics are commonly used in the education sector to assess performance, products, or processes of student learning. Rubrics are gaining importance in organizational performance and program evaluation practice. According to several evaluation practitioners, rubrics can elucidate how excellence and value are defined and applied to evaluation questions or indicators in a given context. This practice note summarizes a pilot project of the National Research Council Canada (NRC) using evaluative rubrics for characterizing relevance and generating conclusions in an evaluation.

Keywords: performance, rubrics, science

INTRODUCTION
Rubrics are a qualitative assessment tool commonly used in the education sector to measure student performance and provide descriptive feedback. They are also now gaining importance in organizational performance and program evaluation practice, where they have been found useful in clarifying and setting out the basis...
on which “judgments about performance, quality, usefulness, and effectiveness are made” (King, McKegg, Oakden, & Wehipeihana, 2013, p. 12). In general terms, rubrics contain two key components: criteria and a scale of performance or quality (i.e., ratings). Evaluative rubrics are commonly represented in a cross-referencing table that describes the performance, value, and effectiveness of a program at different levels of performance based on evidence for a given criterion (Davidson, 2010). According to this approach, rubrics make “transparent how quality and value are defined and applied for each evaluation question or indicator given the context” (Davidson, 2010). When rubrics are well developed and implemented, they can substantially increase the use and credibility of evaluation (King et al., 2013).

The National Research Council (NRC) used rubrics in a recent program evaluation to:

- deliver a clear verdict on program performance based on well-defined performance levels and criteria; and
- create a shared understanding between program managers and the evaluators of the criteria for each performance level.

This practice note describes NRC’s experience piloting evaluative rubrics. It first describes NRC’s experience in developing rubrics followed by a discussion of lessons learned.

**DEVELOPMENT OF RUBRICS**

The steps followed by NRC to develop rubrics were adapted from Davidson, Wehipeihana, and McKegg (2011) and are provided in Table 1.

**Using a collaborative approach**

Rubrics were developed as part of the evaluation planning phase of one of NRC’s research programs, in collaboration with program managers. The evaluation planning phase also included the development of a logic model and evaluation framework, including indicators, through a data and document review as well as key informant interviews. Following this, a 90-minute roundtable discussion was held with program managers to validate the evaluation framework and collaboratively develop rubrics for selected indicators.

The evaluation literature highlights the importance of using a collaborative process in developing rubrics. Collaborative rubric development facilitates stakeholder engagement, and sets expectations and standards that are realistic within the current organizational context (Adams, Nnawulezi, & Vandenberg, 2015; Davidson, 2010). Judy Oakden (2013) also noted that one of the benefits of collaboration is a shared understanding.

Consistent with the literature, NRC found that the roundtable discussion facilitated the program manager’s buy-in to the evaluation and trust in the evaluation

doi: 10.3138/cjpe.31128
Table 1. Steps for the Development of Evaluative Rubrics (Embedded in the Evaluation Planning and Design Process)

Use a collaborative approach
1. Develop logic model and identify evaluation questions through consultations with key program staff.
2. Develop evaluation framework.
3. Coordinate a roundtable session to validate the evaluation framework (evaluation questions and indicators) and develop rubrics for select indicators.
4. Provide roundtable discussion package to participants in advance of the session that includes an agenda, background information on rubrics, evaluation framework and logic model.

Choose the type of rubric
5. Consider whether a holistic or analytical rubric addresses the objectives of the evaluation.

Define performance
6. Moderate the roundtable discussion to:
   a. Validate the evaluation framework.
   b. Discuss the needed or optimal number of performance levels.
   c. Brainstorm what distinguishes excellent, good/adequate, and poor performance (or with more levels as desired).
   d. Draw boundaries around the dimensions that should be included or excluded from the criteria for achieving each level of performance.
   e. Translate these differences and boundaries in distinct dimensions to be articulated in a set of detailed rubrics and ensure that mutually exclusive levels of performance are defined.
   f. If needed, discuss the relative importance or weight of performance dimensions and rubrics associated to each indicator.

Finalize rubric
7. Write the rubric considering the input of key staff members and key literature to help with specific performance concepts or wording, including statistical considerations if quantitative analysis of rubrics scoring is needed.
8. Share the rubrics with roundtable discussion participants for detailed feedback and further debate, as needed.
9. If possible, share the rubrics with a subject matter/industry expert to validate performance criteria. Such expert should be external to the organization and could be a member of an evaluation advisory committee.
10. Revise, adjust and finalize the rubrics.
team. The roundtable discussion was also highly effective in establishing a common language between the program and the evaluation team, providing contextual information that enhanced the evaluation framework and rubrics, as well as building consensus for the rubrics (e.g., what criteria define each level of performance).

**Choosing the type of rubric**

The NRC evaluation team had to choose the type of rubric it would use to meet the objectives of the evaluation. According to the literature, there are two types of rubrics commonly used to assess performance: holistic and analytic rubrics (Brophy, 2012; Danks & Allen, 2014). An analytic rubric presents a description of each level of achievement for each criterion, evaluates multiple criteria along a separate scale or continuum, and provides each criterion with a separate score. Conversely, a holistic rubric consists of a single scale with all criteria included in the evaluation being considered together. Compared to a holistic rubric, an analytic rubric provides more detailed feedback on performance and facilitates a program manager’s ability to make improvements. Consequently, analytic rubrics take a significant amount of time to develop. And while holistic rubrics are quick to use, they are not suitable for complex issues (Brophy, 2012; Danks & Allen, 2014).

For the NRC evaluation, an analytic rubric was used and deemed more appropriate because it afforded additional detailed feedback on the level of performance. Table 2 illustrates two examples of analytic rubrics developed by NRC for relevance indicators.1

**Defining performance**

Following the choice of rubric type, the NRC evaluation team then defined performance. This included identifying the number of performance levels, the criteria that define each level of performance, and the way in which performance is scored. According to Tierney and Simon (2004), a series of six questions needs to be considered when creating rubrics. These questions, which informed NRC’s approach, are presented in Table 3.

**Identifying performance levels and criteria**

The number of performance levels, which are typically divided into three- to six-point scales, can vary from project to project (Center for Institutional Effectiveness, 2011). The most commonly used scales, as found in the literature, are four- or five-point scales including an “insufficient evidence” category. A greater number of performance levels requires increased effort to develop and validate performance descriptors.

NRC used three levels of performance plus an insufficient evidence category. The choice to use a three-point scale was the preference of the NRC program managers and aligned with a three-point scale the program already used to monitor the health of projects. NRC’s approach of developing rubrics in collaboration with the program allowed program managers to take ownership of the rubric development process and to incorporate considerations of organizational culture.
Table 2. Sample Analytic Rubrics from the NRC Pilot Project for Two Relevance Indicators

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Dimension</th>
<th>Excellent</th>
<th>Good/ Adequate</th>
<th>Poor</th>
<th>Insufficient evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence of uniqueness, complementarity, or overlapping of program value proposition (and capabilities) vis-à-vis other S&amp;T organizations in related sectors.</td>
<td>1</td>
<td>Program value proposition and competitive advantage are clearly articulated and effectively communicated to stakeholders (vis-à-vis other S&amp;T organizations in related sectors).</td>
<td>Program value proposition and competitive advantage are somewhat articulated and communicated to stakeholders.</td>
<td>Program value proposition and competitive advantage are not articulated nor communicated to stakeholders.</td>
<td>Information on the level of uniqueness, complementarity, or overlapping of program value proposition (and capabilities) vis-à-vis other S&amp;T organizations in related sectors tended to be unavailable or insufficient to determine performance.</td>
</tr>
<tr>
<td>Program is proactive and forward looking by both managing the current needs and considering future directions/needs of stakeholders to identify new capabilities and facilities needed.</td>
<td>2</td>
<td>Program is focused on the present by managing the current needs in terms of capabilities and infrastructure, but is not considering future directions/needs of stakeholders.</td>
<td>Program is reactive by not managing the current nor considering future needs in terms of capabilities and infrastructure.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program is developing new capabilities (e.g., people) and unique facilities/infrastructures to address future needs.</td>
<td>3</td>
<td>Program is developing capabilities (e.g., people) and facilities/infrastructures to address current needs.</td>
<td>Program is not developing capabilities (e.g., people) and/or facilities/infrastructures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program's capabilities are viewed as quite unique vis-à-vis other S&amp;T organizations in related sectors.</td>
<td>4</td>
<td>Program is viewed as a valued supplier of R&amp;D services and for scientific excellence in technology development in program areas.</td>
<td>Program is viewed as one of the suppliers of R&amp;D services and technology development.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Excellent refers to the highest level of performance, Good/Adequate refers to a satisfactory level of performance, Poor refers to a lower level of performance, and Insufficient evidence refers to a lack of evidence to determine performance.
<table>
<thead>
<tr>
<th>Indicator</th>
<th>Dimension</th>
<th>Excellent</th>
<th>Good/Adequate</th>
<th>Poor</th>
<th>Insufficient evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of complementarity and/or overlapping role and responsibilities of the program vis-à-vis other federal organizations</td>
<td>1</td>
<td>The program is collaborating with all the main federal departments/agencies with mandates that are related to those of the program's areas of activity (as there is no obligation for collaboration). Opportunities for further collaboration are identified and considered.</td>
<td>The program is collaborating with most of the main federal departments/agencies with mandates that are related to those of the program's areas of activity (as there is no obligation for collaboration). Opportunities for collaboration are identified, but not yet considered.</td>
<td>Other federal departments/agencies are not (or avoiding) collaborating with the program by working with other service providers (for services/expertise offered by the program). Opportunities for collaboration are not identified nor considered.</td>
<td>Information on the collaboration activities of the program with federal departments/agencies (and on their activities in the program's areas) tended to be unavailable or insufficient to determine performance.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>The program is viewed by other federal departments/agencies as &quot;the place to go&quot; (e.g., resources, expertise, problem solving, and participation in committees).</td>
<td>The program is viewed by other federal departments/agencies as a valued supplier of R&amp;D services and for scientific excellence in technology development in the program areas.</td>
<td>The program is viewed as one of the federal departments/agencies working in support of the industry sectors related to the program areas.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>The program does not compete with other federal departments/agencies.</td>
<td>The program competes with some activities or services provided by other federal departments/agencies.</td>
<td>The program competes with several other federal departments/agencies.</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Considerations When Creating Rubrics

1. How many performance levels should be used, considering current organizational practices and complexity of the evaluation framework?
2. What distinguishes excellent, good/adequate, and poor performance?
3. Are all the performance dimensions explicitly stated?
4. Are the attributes explicitly stated for each performance dimension and are the underlying characteristics of the performance dimension known and measurable?
5. Are the dimensions equally important (or weighted) for each indicator?
6. Are the boundaries around the dimensions for each criterion, for achieving each level of performance, clear and mutually exclusive?

Adapted from Tierney and Simon (2004)

and processes (i.e., use of a rating scale NRC program managers were familiar and comfortable with). As King et al. (2013) highlight, rubrics are a flexible and adaptable tool, and like good evaluative criteria, tend to be specifically tailored to the evaluation context and its use.

Scoring rubrics

There are several ways to score performance in rubrics. It is possible to assign a score for each criterion used to define an indicator. This approach allows evaluators and program managers to identify areas where the program is doing well and areas that need improvement.

It is also possible to compile an aggregated score for the indicator by taking an average of the scores assigned to each criterion in that indicator. One potential issue with this approach is the representativeness of the score for the indicator (i.e., taking an average of all criteria is not representative of the individual scores). In this regard, the validity and reliability of rubrics may be affected (Moskal & Leydens, 2000). If taking an average score for the indicator, consideration needs to be given to whether each criterion in the indicator is weighted equally. Weighted rubrics are used to reflect that some criteria may be of more significance than others. A common way to weight criteria in the education literature is to assign a percentage to each of them.

In the NRC evaluation, the analytic rubrics were not designed to quantify or attribute a score to each criterion. While the original intent was to weight the different criterion of each indicator, the evaluators decided against it because of the complexity involved in the program (i.e., multiple projects, infrastructure/equipment).

LESSONS LEARNED

Following the NRC pilot project, the evaluation team determined that the development and use of rubrics is not as straightforward as originally expected. In the words of King et al. (2013), “Rubrics, while useful, are not a panacea and they
are not as easy as they may appear. Like any skill, you can learn the theory, but it takes time and experience to become a skilled practitioner (p. 13). As a result of the pilot project, the NRC evaluation team identified several lessons learned. Contributing to the lessons learned were findings from postproject consultations with the program managers involved in the development and use of rubrics. The key advantages and disadvantages of using rubrics stemming from this pilot are summarized in Table 4. The two major lessons learned, related to performance criteria and using rubrics in report writing, are discussed below.

**Performance criteria**

It was generally agreed upon that more time should be allocated in the future for the development of performance criteria to fully benefit from rubrics. While the NRC pilot project made use of additional resources (i.e., approximately two full-time equivalents above what is normally invested in the planning stage of an evaluation), insufficient time affected the quantity (number) of rubrics that could be developed and the quality of performance criteria that were developed. The NRC evaluation team reflected that, in the future, an additional three to four weeks of time would be needed if rubrics were to be effectively developed and used in the analysis and report writing stage.

Also, the NRC evaluation team felt that it would be beneficial to hold two 90-minute roundtable sessions (as opposed to one 90-minute session) to develop and validate the rubrics with program managers. Using this two-stage approach, the first session would address the evaluation framework and the second would be fully dedicated to the development of rubrics. Having one roundtable session focused on the evaluation framework is important because the discussion around language, contextual factors, and performance targets for each indicator of the evaluation framework serves as the foundation for rubrics.

**Quantity of performance criteria**

As a result of insufficient time in the pilot project, performance criteria could not be developed for all indicators. As such, the NRC pilot project was only able to develop rubrics for select relevance indicators, and select efficiency and economy indicators. To develop rubrics for performance-related indicators, additional time would have been necessary to review the available data and define performance targets.

**Quality of performance criteria**

Insufficient time also contributed to several challenges with the quality of performance criteria that were developed. The NRC evaluation team reflected that each rubric contained too many criteria that had to be met to receive a given performance level. For example, to achieve an excellent rating in one case, three criteria needed to be met. However, in cases where one or two of the criteria were met, but not the second or third, it was difficult to assign a rating. The evaluators were left questioning whether meeting one or two criteria warranted an excellent rating as opposed to a lower rating. As such, some degree of subjectivity was
### Table 4. Summary of Key Advantages and Disadvantages of Using Rubrics

<table>
<thead>
<tr>
<th>Section</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of rubrics</td>
<td><strong>Analytic</strong>&lt;br&gt;• help to determine information to be collected (based on criteria)&lt;br&gt;• time consuming to develop&lt;br&gt;&lt;br&gt;<strong>Holistic</strong>&lt;br&gt;• not suitable for complex issues&lt;br&gt;• no detailed feedback&lt;br&gt;• difficult to assign overall score</td>
<td>• help to determine information to be collected (based on criteria)&lt;br&gt;• time consuming to develop&lt;br&gt;• not suitable for complex issues&lt;br&gt;• no detailed feedback&lt;br&gt;• difficult to assign overall score</td>
</tr>
<tr>
<td>Application</td>
<td><strong>Analytic</strong>&lt;br&gt;• more detailed feedback on performance&lt;br&gt;• shared understanding between evaluators and clients&lt;br&gt;• easier for management to make improvements&lt;br&gt;<strong>Holistic</strong>&lt;br&gt;• quick to use&lt;br&gt;• useful for single dimension criteria</td>
<td>• demands a lot of effort and experience&lt;br&gt;• collaboration is needed for the development&lt;br&gt;• evaluation becomes limited by rubrics criteria</td>
</tr>
<tr>
<td>Performance levels</td>
<td>• criteria are more focused and concise&lt;br&gt;• ease of use&lt;br&gt;• many different uses&lt;br&gt;• a flexible tool&lt;br&gt;• collaboration is needed for the development&lt;br&gt;• more credible results&lt;br&gt;• more useful for the client</td>
<td>• more levels used, the more effort needed in development&lt;br&gt;• difficult developing score&lt;br&gt;• difficulties in assigning overall score criteria</td>
</tr>
<tr>
<td>Scoring</td>
<td>• possibility of an average indicator (weighting sections)&lt;br&gt;• a numerical value can be assigned to each level&lt;br&gt;• provides consistency in scoring</td>
<td>• time-consuming process&lt;br&gt;• large time investment from client</td>
</tr>
<tr>
<td>Collaboration</td>
<td>• shared understanding</td>
<td>• time-consuming process&lt;br&gt;• large time investment from client</td>
</tr>
</tbody>
</table>
introduced. Postproject consultations with program managers also revealed that multidimensional criteria made it difficult for the program to validate their overall performance score. The difficulty with the rating criteria appears to be common, as Stone-Jovicich (2015) reported similar challenges. In the future, criteria used to define each level of performance should be distinct. As clarity and distinction are important to client comprehension, the criteria should not contain too many ideas.

The program managers also noted that it would have been a good practice to have external experts in their field consulted to ensure that the definitions of the performance criteria were appropriate and accurately represented the standard expected for the level of performance. This step may help ensure the quality of the performance criteria.

**Report writing**

One of the reasons that the NRC evaluation team first sought to pilot evaluation rubrics was for the oft-cited benefit of rubrics simplifying and expediting report writing. This, however, was not the experience of the NRC evaluators during the pilot and is a lesson learned. The NRC evaluation team found that in practice, rubrics made report writing harder because findings were bound by predetermined criteria. More specifically, rubrics hindered overall findings as they were not able to capture how different aspects interacted and what emerged as a result of those interactions. To some extent, the development of evaluation findings was constrained by the language of the rubrics. The challenge was reporting on the findings according to the rubrics while still being able to integrate additional findings. Integrating rubrics into the report also proved difficult, as not all key aspects of performance explored in the evaluation were outlined in the rubrics.

As a result of these challenges, NRC did not include the rubrics in the final evaluation report. The challenge posed by rubrics in facilitating emergent findings was also highlighted by Stone-Jovicich (2015). Ultimately, in the NRC pilot project rubrics ended up being an analytical tool that helped the evaluators identify data needs, collect and organize data, and make judgements about performance; they were not a reporting tool. Future use of rubrics will benefit from realistic expectations about their role and where they can add value to the evaluation process.

**CONCLUSION**

As a result of this pilot project, the NRC evaluation team reflected that it is unsure about the feasibility and practicality of using rubrics in NRC evaluations given the challenges experienced. That said, there were many benefits, particularly for stakeholders, as a result of the rubrics (e.g., client’s understanding of what was being assessed was enhanced, evaluators’ understanding of the program was improved). Due to these benefits, NRC will employ rubrics in a subsequent evaluation, incorporating the lessons learned from this first pilot project, before rendering a judgement as to their viability as an effective evaluation tool.
Based on the NRC evaluation teams experience, key actions for success include working with invested evaluation clients, having distinct and clear criteria for the rubrics, and defining a scale that is understood by all parties involved. In addition, realistic expectations about the role rubrics can play in the evaluation process will facilitate the successful use of rubrics in evaluations (e.g., as a data collection and analytical tool as opposed to a reporting tool).

ACKNOWLEDGEMENTS
We would like to thank the program managers for taking part in this pilot project as well as taking the time to provide feedback for the practice note. We would also like to acknowledge Michaela McConnell, former Student Auditor at NRC, for her support in the drafting of the practice note, and Julian King (Julian King & Associates Limited) and Judy Oakden (Pragmatica Limited) for sharing presentations and publications on their experience using evaluative rubrics.

NOTES
1 Readers interested in the full set of rubrics developed for the NRC pilot project are asked to contact the lead author for a copy.

REFERENCES


**AUTHOR INFORMATION**

Ghislaine Tremblay is Director, Evaluation at the National Research Council. She previously held the position of Director of Evaluation at the Canada Foundation for Innovation. She has significant experience in conducting science and technology evaluations and in the management of the evaluation function. Ghislaine holds a Graduate Diploma in Program Evaluation and is a Credentialed Evaluator.

Frédéric Bertrand is an independent senior evaluation consultant (Ber /). Over the last 15 years, he has collaborated with more than 60 science and technology organizations, universities, foundations, and private firms by supporting managers and internal evaluators, designing and managing applied studies, and leading evaluations. Frédéric is a Credentialed Evaluator, holds a Masters in Environmental Sciences, and regularly supports the Canadian Evaluation Society (CES) for sponsorship activities and the student case competition.

Melissa Fraser is a Senior Evaluation Officer at the National Research Council. Melissa has ten years of experience evaluating science and technology programs. Methodologically, Melissa has significant experience with both qualitative and quantitative research methods, and holds a specialization in inferential statistics.
Abstract: The terms cultural responsiveness and cultural competence have become ubiquitous in many fields of social inquiry, including in evaluation. The discourse surrounding these issues in evaluation has also increased markedly in recent years, and the terms can now be found in many RFPs and government-based evaluation descriptions. We have found that novice evaluators are able to engage culturally responsive approaches to evaluation at the conceptual level, but are unable to translate theoretical constructs into practice. In this article we share a framework for teaching culturally responsive approaches to evaluation. The framework includes two domains: conceptual and methodological, each with two interconnected dimensions. The dimensions of the conceptual domain include locating self and social inquiry as a cultural product. The dimensions of the methodological domain include formal and informal applications in evaluation practice. Each of the dimensions are linked to multiple domains within the Competencies for Canadian Evaluation practice. We discuss each and provide suggestions for activities that align with each of the dimensions.

Keywords: culture, cultural competence, cultural responsiveness, teaching evaluation

Résumé : Les termes sensibilité culturelle et compétence culturelle sont maintenant omniprésents dans de nombreux domaines d’enquête sociale, notamment en évaluation. Le discours entourant ces questions en évaluation s’est aussi intensifié de façon marquée au cours des dernières années et ces termes sont maintenant présents dans de nombreuses demandes de proposition et descriptions d’évaluation émanant d’organismes gouvernementaux. Nous avons trouvé que les évaluateurs débutants sont en mesure de concevoir des approches d’évaluation culturellement adaptées, mais sont incapables de transférer ces notions théoriques à la pratique. Dans le présent article, nous décrivons un cadre pour l’enseignement d’approches évaluatives qui soient culturellement sensibles. Le cadre inclut deux sphères – conceptuelle et méthodologique – chacune ayant deux dimensions interconnectées. Les dimensions de la sphère conceptuelle implique de positionner l’évaluateur et le processus de recherche comme un produit culturel. Les dimensions de la sphère méthodologique comprend des applications formelles et informelles pour la pratique

Corresponding author: Ayesha S. Boyce, Educational Research Methodology, School of Education, University of North Carolina at Greensboro, ayesha.boyce@uncg.edu

© 2017 Canadian Journal of Program Evaluation / La Revue canadienne d’évaluation de programme
32.2 (Fall/automne), 266–279 doi: 10.3138/cjpe.31132
Teaching Culturally Responsive Approaches to Evaluation

Mots clés : culture, compétence culturelle, sensibilité culturelle, enseignement de l'évaluation

INTRODUCTION

The terms cultural responsiveness and cultural competence have become ubiquitous in many fields of social inquiry, including in program evaluation. Simply put, culturally responsive evaluation has been defined as responsive evaluative inquiry that meaningfully attends to and addresses the cultural context of the community (SenGupta, Hopson, & Thompson-Robinson, 2004). Evaluations can achieve cultural competence by being responsive to the needs of the program community and specifically tailored to the unique groups and communities of focus (Hopson, 2009). Culture is dynamic and ever-changing, and as such, evaluators are encouraged to continuously seek to understand the culture, context, historical perspective, power, oppressions, and privilege in each new evaluation context (Greene, 2005; Pon, 2009; Symonette, 2004). With justification rooted in our field’s continued expansion in racially, ethnically, linguistically, economically, politically, and culturally diverse international contexts (Hood, Hopson, & Frierson, 2015), many global voluntary organizations for professional evaluation (VOPEs) have developed public statements related to culture, cultural competency, and cultural responsiveness (Kosheleva & Segone, 2013). In addition, training of culturally responsive and equity-focused approaches are offered within university courses (Davies & MacKay, 2014) and at VOPE conferences around the world (Catsambas, Segone, de Silva, & Saunders, 2013). The discourse surrounding these issues in evaluation has also increased markedly in recent years, and the terms can now be found in many Request for Proposals (RFPs) and government-based evaluation descriptions (Botcheva, Shih, & Huffman, 2009).

While a definitive relationship between evaluation theory and practice remains elusive (Christie, 2003; Cooksy, Mark, & Trochim, 2009; Tourmen, 2009) and in need of further study (Chelimsky, 2013), our initial examinations indicate that attempts at cultural responsiveness in practice are in need of attention. As instructors of six evaluation courses within an educational research methodology department with program evaluation–focused M.S. and Ph.D. program tracks, we endeavour to train evaluators who not only have strong technical and methodological skills, but who also think well and critically about how to meaningfully and responsively attend to culture and context within an evaluation setting. As with other theoretical evaluation approaches, we have found that novice evaluators, students, and practitioners are able to engage culturally responsive approaches to evaluation at a theoretical level, but remain challenged when translating their theoretical constructs into practice (Chouinard & Boyce, in press, a; Chouinard et al., 2016). For example, novices have trouble understanding how to design...
data collection instruments that are culturally commensurate, how their personal biases influence the evaluation, and how issues of race, power, inequity, diversity, and culture might influence their relationships with stakeholders. Moreover, our research, as well as our observations as reviewers of journal articles, conference proposals, and evaluation plans, would indicate that, in many cases, attempts at culturally responsive evaluation are little more than the inclusion of symbolic and politically correct buzzwords (Chouinard & Cousins, 2009). As the conversation surrounding culture, cultural competence, and cultural responsiveness is increasing exponentially, the need to teach novice evaluators both the theory and practice of culturally responsive approaches becomes critical. In this article we share a framework that we have used for teaching culturally responsive approaches to evaluation, followed by implications for evaluation practice.

**DIMENSION OF PRACTICE**

Ideally, all evaluators should attend to the social, ethical, political, cultural, and value dimensions of any evaluation context (Chouinard, 2016), and aim for nuanced and responsible methods to evaluation (Hood, Hopson, & Frierson, 2015). As such, several evaluation approaches have surfaced to guide evaluators in meaningfully and responsively engaging in diverse contexts, with what Kirkhart (2005, 2010) refers to as “multicultural validity.” Evaluation frameworks that guide practitioners to explicitly address issues of power, social justice, equity, human rights, and cultural complexity include transformative participatory (Cousins & Whitmore, 1998), transformative (Mertens, 1999, 2009), democratic (Kushner, 2005; MacDonald, 1976), deliberative democratic (House & Howe, 2000), equity focused (Segone, 2011), critical evaluation (Everitt, 1996; Fay, 1987), values-engaged, educative (Greene, Boyce, & Ahn, 2011; Greene, DeStefano, Burgon, & Hall, 2006), restorative justice (Chouinard & Boyce, in press, b), and cultural and contextually responsive approaches (Frierson, Hood, Hughes, & Thomas, 2010; Hopson, 2009; Madison, 1992; Thomas & Stevens, 2004). What seems clear is that more evaluators have been intentional about anchoring their work in inclusive, democratic, and culturally responsive ideals (Frierson et al., 2010; Greene, 2006; Mertens & Wilson, 2012). With more than 200 articles that mention culturally responsive or culturally competent evaluation in the literature (Hood, Hopson, & Frierson, 2015), empirical examinations of this phenomenon have also gained significant traction (see Chouinard & Cousins, 2009; Chouinard & Hopson, 2016; Samuels & Ryan, 2011).

No program or evaluation is value-free or culture-free (House, 1980). Although multiple definitions exist, culture is generally thought of as representing the shared norms, values, and assumptions of a group (Samuels & Ryan, 2011; SenGupta et al., 2004). Culture can also refer to shared language, gender, ethnicity, race, religion, social class, sexual orientation, disability, age, and/or geographic location (Merriam-Webster, 2011; Bowen & Tillman, 2015). Theoretically, culturally responsive evaluation is situated at the intersection of (a) decolonizing/indigenous epistemologies and frameworks, (b) critical theories of epistemologies and race, and (c) social agenda and advocacy theories and approaches in evaluation.
(Hopson, 2009). In culturally responsive practice, evaluators are encouraged to analyze the program’s cultural and sociopolitical context, develop evaluation questions with the active inclusion of multiple stakeholders, utilize culturally commensurate data collection and analyses methods, and share findings with a variety of audiences (Frierson et al., 2010). Culturally responsive evaluators generally aim to be collaborative, respectful, and attentive, and to honour cultural norms, illuminate structural injustices, promote action to redress inequities, and be reflective about their own culture, prejudices, assumptions, and biases.

METHODS

We are both tenure-track assistant professors within an educational research methodology department. In addition to teaching other graduate methodology courses, we also teach an introduction to evaluation, a practicum in evaluation, advanced evaluation theory, public policy and evaluation, collaborative approaches, and a culturally responsive approaches to evaluation course. We are both committed to developing students’ technical skills needed for evaluation practice, while also encouraging the adoption of a critical lens (Everitt, 1996; Fay, 1987) and attention to social justice. As Thomas and Madison (2010) have argued, “evaluation students also must be inspired to challenge the status quo, to care about the interests of the disadvantaged, and to uncover weaknesses within the system that contribute to inequities within society” (p. 571). We both have a social justice orientation that is reflected in the courses we teach, our research interests, and our evaluation practice. Our hope is that students who complete our courses will have a strong methodological foundation, and conduct evaluations that are based in democratic principles and that promote equity, fairness, and diversity.

Our insights for this article are based on reflections from a graduate course in culturally responsive approaches to research and evaluation that we co-taught during the Spring 2016 semester. The aim of this course was to provide a comprehensive overview of culture and its centrality in evaluation and research practice. The focus of the course was on culturally responsive approaches to evaluation and research in educational settings, with a combination of both theoretical and practical applications. The course was conducted as a mix of lecture, student-led seminars, and group learning activities. The course provided an overview of approaches that are considered culturally responsive, and many of the readings examined key dimensions of practice (e.g., context, relationships, validity, methodology, and design). Students were required to keep a journal throughout the semester, and were encouraged to reflect personally and academically on the course readings and classroom experiences. In their journals they highlighted theoretical wrestlings, noted interpretive insights, and attempted to define their own cultural location. Aside from ongoing class discussions, other formal course assignments included a position paper on an issue or dilemma confronting researchers working in culturally diverse communities, and a final paper integrating culturally responsive theory and practice with their own research interests.
Prior to the start of the semester, we decided to systematically capture how the course was planned, designed, implemented, and evaluated with particular attention to logistics, pedagogy, and strategies. As course instructors, we met weekly throughout the semester to critically reflect upon our pedagogy and student learning, share resources, and offer constructive feedback as we prepared for the following week. Each week we took notes of the main ideas and topics covered during these reflections. At the end of the semester, we reviewed our reflective notes as a way to better understand course triumphs and challenges, and to map out changes for future iterations of the course.

LESSONS LEARNED AND IMPLICATIONS FOR PRACTICE

Our experiences teaching this course demonstrate that while students were able to understand cultural responsiveness intellectually, they were unable to easily translate their concepts into practice. For example, while some students were able to understand ongoing issues of trustworthiness within historically marginalized populations, they were unable to identify what this would mean when working within these communities, how they would address responsiveness to stakeholder needs, and how they would include voices that represented diverse stakeholder perspectives (Bowman, Dodge Francis, & Tyndall, 2015; LaFrance, 2004; Smith, 1999). Others were able to recognize the need to develop culturally and contextually appropriate data collection instruments for specific populations, but were unable to move beyond the suggestion that this could not be accomplished merely through simple language translation.

To aid in puzzling through these experiences, we developed a conceptual framework to help us better understand pedagogy as it related to teaching culturally responsive approaches to evaluation. The framework depicted in Figure 1 includes two domains: conceptual and methodological, each with two interconnected dimensions. The dimensions of the conceptual domain include two foci: locating self and social inquiry as a cultural product. The dimensions of the methodological domain are formal and informal applications in evaluation practice. Each of the dimensions are linked to multiple domains within the Competencies for Canadian Evaluation practice (Canadian Evaluation Society, 2010; Stevahn, King, Ghere, & Minnema, 2005). The conceptual dimension is connected to the reflective, situational, and interpersonal competency domains. The methodological dimension is connected to the technical, situational, management, and interpersonal practice competency domain. Ideally, teaching would occur at the intersection of all dimensions, and include instruction and activities across both domains. What follows is a discussion of each dimension, with suggested activities aligned to each dimension.

Locating self

The lens through which we view the world influences all evaluation processes from design to implementation and interpretations (Milner, 2007; Symonette,
Figure 1. Conceptual framework for understanding pedagogy as it relates to teaching culturally responsive approaches to evaluation.
To think well about issues of culture, power, equity, class, race, and diversity, evaluators should first seek to understand their own personal and cultural values, all of which are influenced by their educational backgrounds, personal identities, experiences, underlying values, prejudices, predispositions, beliefs, and intuition (Symonette, 2004). As Peshkin (1988) has noted, the practice of locating oneself can result in a better understanding of one's own subjectivities.

To advance our students' understanding of their own cultural locations, they kept a reflective journal throughout the semester. At the beginning of the course, students were asked to reflect on their own backgrounds and cultural locations, with a focus on their own social identity, the identity of others, whether they belong to any groups with power and privilege, and how their educational background and identities shape their role and experiences (personal, societal, and research). We then invited students on a “privilege walk” that began with all students standing in a line together, shoulder to shoulder. We had several prompts and, based on responses, students took a step forward, a step backward, or stood still. A few examples of these prompts were: if you are a white male, take a step forward; if your work holidays coincide with religious holidays you celebrate, take a step forward; if you have visible or invisible disabilities, take a step back; if you took out student loans to advance your education, take one step back; if you attended private school or summer camp, take one step forward. At the end of the activity, students engaged in a discussion of their final locations in comparison to the locations of their peers. Further class discussion also included student reflections on their own values in response to current events. For example, we watched the Beyoncé video *Formation* and asked students to share thoughts and reactions, after which they were asked to reflect upon the underlying values associated with their reactions. If they were offended, they shared what specifically about the video they found offensive. If they felt vindicated or connected with the video, they offered reasons why.

**Social inquiry as a cultural product**

It is imperative that evaluators understand individual, institutional, societal, and civilizational racism and its intersections with knowledge production in the social sciences (Scheurich & Young, 2002). Cultural responsiveness, sensitivity, and competence was born out of the need to dismantle archaic discourses of power and inequity. As Symonette (2004) proclaims, “evaluators need to proactively interrupt the operation of critical autoload of default settings because they result in trust-eroding inaccuracies, truncated understanding, and twisted representations” (p. 97). Historically, “communicentric bias”—the tendency to make one's own community, often the majority class, the centre of conceptual frames that constrains all thought—has resulted in negative consequences for minority populations (Gordon, Miller, & Rollock, 1990). Further, social science knowledge of minority populations has demeaned characteristics, distorted interpretations of conditions and potential, and remained limited in its capacity to inform efforts to understand and improve life chances of historically disadvantaged populations.
(Johnson et al., 2008; Ladson-Billings, 2000). While efforts to dismantle communicentric bias and epistemological racism have been underway for more than 80 years (Hood, 2001), knowledge production in and of itself is still a cultural product. To engage with these topics, we began with conversations about social science paradigms and discussed the role of ontology, epistemology, axiology, and methodology in social inquiry. Students were required to read, critique, and discuss articles and chapters related to epistemologies of difference, racialized discourses, and critiques about the nature of social inquiry.

**Formal applications in practice**

Numerous scholars have addressed the implications of cultural responsiveness in practice (Frierson et al., 2010; Hood, Hopson, & Kirkhart, 2015), with some encouraging contemplation surrounding threats to, as well as evidence for, multicultural validity by examining relational, consequential, theoretical, experiential, and methodological justificatory perspectives (Kirkhart, 2005, 2010). However, as previously mentioned, we have found that novice evaluators need additional practical examples, supplemental guidance, and multiple iterations to work through what the application of these approaches look like in practice. As such, we engaged students in activity-based practice to provide them with the opportunity to apply cultural responsiveness in planning, designing, and implementing an evaluation.

As an aid, we developed three case study contexts and asked students to work through formal methodological applications. We purposely chose racially, ethnically, linguistically, economically, politically, and culturally diverse global contexts. These included an evaluation of a government program for Quilombos, descendants of an African slave community in Brazil; an evaluation of a dating violence prevention program for American Indian youth; and a research project to gain knowledge about family violence within immigrant and refugee communities. We used these case studies to provide students with opportunities to develop evaluation designs that are culturally and contextually appropriate. Students worked within constraints of program contexts to identify key stakeholders and the evaluator’s role, develop evaluation or research purposes and questions, establish strategies to address practical and method-based problems, and consider innovative reporting approaches. After observing group discussions, reading final papers, and reflecting upon the course, we realized that there was a need to also facilitate attendance to potential informal applications of culturally responsive approaches.

**Informal applications in practice**

Evaluation is not simply composed of question development, data collection and analyses, and reporting. While evidence of successful attention and responsiveness to culture should be found in formal evaluation documents, engagement with these issues can also take place through the less formal occasion of dialogue and discussion with stakeholders (Boyce, 2017; Tillman, 2014). Tactful engagement
with sensitive issues, informal actions, and interactions with cultural responsiveness can take many forms. These include bringing up issues of race, power, inequity, diversity, and culture for dialogue in meetings, emails, and conversations with clients, funders, and stakeholders. Evaluators who are committed to social justice will acknowledge differing stakeholder opinions, while also attempting to nudge stakeholders and assist them in surfacing their own values, prejudices, and subjectivities (Greene et al., 2011). As such, evaluations require interpersonal, facilitation, negotiation, and collaboration skills.

To do this we recommend having students and novice evaluators work to become comfortable discussing these topics. To be culturally responsive and to engage with issues of power, values, culture, diversity, and inequity, evaluators need to have access, credibility, and authority within the context (Greene et al., 2011). We believe that case study exercises will assist students in practicing to enlist culturally responsive approaches during informal settings (Patton & Patrizi, 2005). In the next iteration of our course we plan to have students do a small-scale, culturally responsive study so they can practice some of the aforementioned topics.

CONCLUSION

In this article we have provided a framework for teaching culturally responsive approaches to evaluation. The framework includes two domains: the conceptual ground upon which we as evaluators stand, including our own experiential values and culture and the epistemological location on which our practice sits; and the methodological ground upon which we stand, including formal and informal applications in practice. To engage with each of these dimensions, much of the work to be done is reflective and case study–based. We recognize that no evaluator can ever become truly culturally competent, as each evaluation context comprises varying social, ethical, political, cultural, and value dimensions (Symonette, 2004). While we continue to grapple with the development and implementation of culturally and contextually responsive evaluation in our own practice, it is our hope that more and more evaluators will be encouraged to be respectful, be attentive to, and honour cultural norms; illuminate structural injustices; be reflective about their own culture, prejudices, assumptions, and biases; and formally and informally be culturally responsive in their evaluation practice.

REFERENCES


Teaching Culturally Responsive Approaches to Evaluation


Tillman, A. S. (2014). Using a values-engaged educative evaluation approach to address issues of diversity and equity in a multi-site science technology engineering
and mathematics program evaluation (Doctoral dissertation, University of Illinois at Urbana–Champaign). Retrieved from https://www.ideals.illinois.edu/handle/2142/72872


**AUTHOR INFORMATION**

**Ayesha Boyce** is an assistant professor within the Educational Research Methodology Department at the University of North Carolina at Greensboro (UNCG). Her research focuses on attending to value stances and issues related to diversity, equity, access, climate, cultural responsiveness, and restorative justice within evaluation—especially multisite, STEM, and contexts with historically marginalized populations. She also examines teaching, training, and novice learning in evaluation. Dr. Boyce has evaluated many programs funded by the National Science Foundation, the National Institutes of Health, Arizona Department of Education, Title VI, Illinois P-20 Council, and Spencer and Teagle foundations. She currently sits on the American Evaluation Association (AEA) Thought Leaders Committee for “An AEA Dialogue on Race and Class.” From 2014 to 2016 she was the chair of the AEA STEM Education and Training TIG. She teaches classes in program evaluation and research methodology where she emphasizes good practice with mindful attentiveness to theoretical roots.

**Jill Anne Chouinard** is an assistant professor within the Educational Research Methodology Department at the University of North Carolina at Greensboro (UNCG). Her main interests are in cross-cultural/culturally responsive approaches to research and evaluation, participatory research and evaluation, and evaluation and public policy. Dr. Chouinard has extensive experience working on evaluations at the community level in the areas of education and training, social services, health, and organizational learning and change. Much of her evaluation work has been conducted in First Nations and Inuit communities, as well as in other culturally diverse community settings. She maintains a strong connection between her evaluation practice and her academic research. As such, her research and academic writing focuses on participatory approaches to evaluation in culturally diverse community settings, from both a national and an international perspective.
The Impact of Practice on Pedagogy: Reflections of Novice Evaluation Teachers

Jill Anne Chouinard and Ayesha S. Boyce
University of North Carolina at Greensboro

Abstract: In this practice note two novice evaluation teachers share their findings from research conducted with students who were enrolled in a theory and practicum course in evaluation. The study focused on understanding how and in what ways students navigate between the world of theory and the world of practice. The findings from this study subsequently led to a re-envisioning of the course offerings to provide a more nuanced transition between two dichotomized conceptualizations of evaluation (theory and practice), revised syllabi, and the addition of a third course. The implications of this research (and subsequent pedagogical revisions) raise important issues for evaluation teachers and practitioners, as we continue to debate the relationship between theory and practice in evaluation.

Keywords: evaluation teaching, novice evaluators, theory and practice

Résumé : Dans cette note de pratique, deux nouveaux professeurs d'évaluation font part des résultats d'une étude effectuée auprès d'étudiants inscrits à un cours théorique et à un stage en évaluation. L'étude visait à comprendre comment et de quelle façon les étudiants font la transition entre la théorie et la pratique. Les résultats de cette l'étude ont par la suite mené à une refonte de cours pour faciliter la transition entre des conceptualisations dichotomisées de l'évaluation (théorie et pratique), une révision du curriculum et l'ajout d'un troisième cours. Cette recherche (les révisions pédagogiques qu'elle a entraînées) soulève des questions importantes pour les professeurs et les praticiens de l'évaluation, alors que se poursuit le débat sur le lien entre la théorie et la pratique en évaluation.

Mots clés : enseignement de l'évaluation, nouveaux évaluateurs, théorie et pratique

INTRODUCTION

In evaluation, the relationship between theory and practice remains a significant and ongoing concern. As an applied social science, program evaluation involves far more than the simple application of theory (Fitzpatrick, Christie, & Mark, 2009), as the world of practice is often complex, dynamic, and indeterminate, not at all amenable to the direct application of theoretical knowledge (Polkinghorne, 1992; Schwandt, 2015). As evaluation is an interdisciplinary practice, there are a plethora
of competing theoretical perspectives available to assist evaluators in understanding the complexity of the social world and in rendering judgements about programs and policies. As new professors of evaluation who teach an evaluation theory course followed by a practicum course in evaluation, we have observed that our students experienced multiple challenges (and frustrations) as they attempted to navigate the relationship between theory and practice, between what they learned in the classroom and their experiences in the field. Simply put, our students wrestled with the fact that their theoretical understanding of evaluation could not provide them with a step-by-step guide to practice; they felt that the theory course did not adequately prepare them for the dynamic and sociopolitical nature of the evaluation context. As one of the students lamented, “There’s no theory for when everything falls apart.”

To help us better understand the experiences of our students and to help inform our teaching practice, we conducted a qualitative research study at the end of the academic year to further explore the perspectives of five graduate students who had been enrolled in both courses, and who volunteered to be part of our study as participants and coauthors (Chouinard et al., in press). Our research focused on understanding how and in what ways novice practitioners navigate between the world of theory and the “rough ground of practice” (Schwandt, 2003).

The introductory course provided students with an overview of the basic concepts, issues, theories, and models in evaluation. The goals of the course were to enable students to compare and contrast major theories of program evaluation, with a focus on determining which evaluation approach to use in a given context; developing an evaluation plan; identifying the various roles of the evaluator; understanding the social, historical, political, and cultural dynamics of an evaluation context; planning the collection of data; and establishing reporting processes.

The practicum course introduced students to the practical craft of program evaluation, particularly to the practices or methods that characterize evaluation in the field of education. Students planned and conducted a small-scale evaluation, with activities that included evaluation planning, data collection, analysis, interpretation of findings, and ongoing and final reporting processes. The objective of the course was to provide students with first-hand experience in conducting and managing a small-scale evaluation, applying basic data collection tools and using qualitative and quantitative data analysis, and completing both conventional and innovative reporting of results. As part of the practicum course, students were required to keep reflective journals where they were encouraged to record their evaluation experiences, reflections as novice evaluators, and share theoretical wrestlings, interpretive insights, and other challenges.

Our data for this research were based on (a) an analysis of the reflective journals students kept throughout the practicum course, (b) one-on-one interviews with each of the five students based on open-ended questions identified through our review of the literature and our reading of their reflective journals, and (c) a follow-up focus group with students to share our initial analysis and test out
hunches and hypothesis. Our goal throughout was to deepen our understanding of their experiences as novice evaluators in the field, so as to ultimately inform our future evaluation curriculum and teaching practices.

Our study highlighted the three interconnected dimensions of practice that include context, theory, and personal and cultural biographies, all of which dynamically intersect in evaluation practice. These three dimensions shift the focus from a theory-practice binary to a focus on the sociocultural contexts and biographical forces that interact together in shaping evaluation practice. In what follows, we briefly share our key findings from this study and reflect on implications for our current teaching of both evaluation theory and practice.

**STUDY FINDINGS: DIMENSIONS OF PRACTICE**

The focus of our recent study was on understanding the relationship between theory and practice based on the perspectives and experiences of student evaluators, as they moved from the classroom to an engagement with practice. While there has been prior research related to the training of evaluators, the focus of this literature was not explicitly on understanding the relationship between theory and practice (Trevisan, 2004), but on specific aspects of training. For example, Hurley, Renger, and Brunk (2005) focused their research on teaching evaluation in the classroom and in the field; Alkin and Christie (2002) looked at the pedagogical benefits of role-play in training evaluators; Altschuld (1995) and others (Darabi, 2002; Morris, 1994) focused on the design of training courses; Lee, Wallace, and Alkin (2007) looked at a problem-based learning approach to teaching evaluation; Kelly and Kaczynski (2008) explored the use of experiential learning with evaluation novices; Jewiss and Clark-Keefe (2007) examined the use of self-reflection for novices; and Levin-Rozalis and Rosenstein (2003) looked at the effects on mentoring on learning evaluation.

In our study, we privileged the concepts of theory and practice with novice evaluators to shed light on what remains a problematic and persistent dichotomy in evaluation, especially for those who are new to the field. In the social sciences, a tightly coupled relationship between theory and practice endures, as practice is thought to represent the technical expression or application of theory, an idea whose genesis can be traced to the conflation of theory in the natural sciences with theory in the social sciences (Flyvbjerg, 2001). However, theories in the social sciences must be distinguished from theories in the natural sciences, as they cannot provide evaluators with predictive certainty, but rather with what Schwandt (2014) refers to as “aids to the evaluation imagination” (p. 234). Our findings clearly show that the challenges evaluators confront in the field are complex, indeterminate, and highly specific to the sociopolitical and cultural context of the program and community. In other words, they cannot be solved through the simple application of the right theoretical principles. As one of the students described:
There were certain aspects of theory that stuck with us but at the end of the day it goes out the window. It’s kind of like scaffolding. If you’re on a tightrope, it’s the trampoline that’s below you. If you get stuck you can rely on it but it may not get you from one end to the other.

As our students quickly learned, the program and evaluation context is far too dynamic and complex for any one theory to map on perfectly (van Manen, 1995). A key finding in our study was that students’ reification of theory, while likely related to their lack of evaluation expertise and confidence (Flyvbjerg, 2001), paradoxically also made it more challenging for them to let go of their expectation that theory alone could help them address the challenges of the context. Ironically, the messier the context, the greater the need students had for theory to ultimately prevail. The tension that students experienced between their understanding of theory and their actual field experience suggests that there is far more involved than the simple application of technical knowledge (Schwandt, 2003). Evaluation, as our students soon came to realize, is a sociopolitical process (Greene, 2000) that requires situational sophistication beyond the simple application of methodological rigour and skill. Our students’ initial understanding of theory and practice, situated as it was in the intellectual tradition of the natural sciences (Carr & Kemmis, 1986), ultimately could not provide the guidance nor the technical direction they required for an engagement with the exigencies of evaluation practice. As one of the students described, “we were kind of thinking that we would follow a little bit more of a roadmap, instead of building the plane while it’s flying.”

**IMPLICATION FOR PRACTICE**

As new teachers of evaluation, we primarily focused in our study on understanding how students navigate between the world of theory and the world of practice, as a way to inform our teaching practice. After conducting the study (Chouinard et al., in press), we spent the following year revising our course sequence and adapting our syllabi to better reflect our understanding of student engagement with theory and practice. At this point, we have re-envisioned the introductory evaluation course so that it aligns better with the evaluation practicum in the second semester, away from a view of evaluation demarcated by a theory-practice binary. We have also added a third, advanced course in evaluation that is focused almost exclusively on the study of evaluation theory. The goal in our redesign is to ensure that students have a basic understanding of the diversity of evaluation approaches and paradigmatic implications and methods of practice (e.g., evaluation design, logic models, frameworks) before they go out in the field to conduct an evaluation. The advanced theory course in evaluation is intended for students who wish to pursue careers in evaluation. In what follows, we describe our pedagogic modifications as we continue to reflect on how best to teach novice evaluators the theory and practice of evaluation.

*Introductory Course.* While students are introduced to some evaluation theory (e.g., different approaches and theorists), different paradigms, and
methodological implications in this course, the predominant focus is on the technical aspects of evaluation (e.g., constructing a logic model, creating an evaluation plan and budget, developing evaluation questions, methods of data collection and analysis, approaches to reporting). Each class session provides students with hands-on activities related to specific technical skills, with two class sessions now devoted to constructing logic models. We also use case studies (see Patton & Patrizi, 2005) in each class session to provide students with the opportunity to identify appropriate evaluation designs that are culturally and contextually appropriate; understand (and work through) the exigencies of program contexts; identify key stakeholders, evaluation purpose, and the role of the evaluator; develop strategies to address practical and method-based problems; and engage with diverse evaluation orientations. The case study approach also provides students with the opportunity to practice evaluation in a low-stakes setting, to make mistakes and adjustments, to practice reading diverse cultural contexts, to try on different evaluator roles, and to use different data collection methods and approaches based on identified evaluation questions. Our overall goal in this course is to ensure that students have ample opportunity to think through being evaluators and conducting evaluations across a diversity of program and community settings before they ever go out into the field. As Patton and Patrizi (2005) have stated, the use of the case study method enables novice evaluators to “bridge the gap between knowing and doing” (p. 98), a gap that our students have struggled to overcome.

Practicum Course. In this course, students transition from the safety of evaluation practice in the classroom setting to the “real world” of evaluation out in the field, from a case study model to an engagement with a real case. We aim to dissuade students from the expectation that they are methodological technicians whose job is merely to design and implement an evaluation regardless of multiple and very often competing contextual demands. While the study of case studies in the introductory course is not an engagement with a “real” evaluation, it nonetheless provides ample opportunity for students to learn about evaluation across multiple, diverse community and cultural program settings. In the practicum course, students are divided into small teams of two to three students, and together they select a program to evaluate among programs (and organizations) that we have preselected. Students work directly with their clients to design and conduct a small-scale evaluation, work that includes preparing an evaluation plan, collecting data, analyzing data, interpreting findings, and reporting. While the majority of class meetings are discussion-based, with dialogue about issues, problems, and solutions constituting the majority of the class session, the first few weeks are designed as workshops where students learn how to create an evaluation plan (evaluation design, logic model, evaluation framework, etc.). The overall goal of this course is to provide students with first-hand experience in conducting and managing a small-scale evaluation, applying basic data collection tools and using qualitative and quantitative data analysis, and completing both conventional and innovative reporting of results.
Advanced Evaluation Theory Course. This new, advanced course provides students with the opportunity to critically examine a diversity of evaluation approaches and models from different evaluation roots (Alkin, 2013; Mertens & Wilson, 2012). The course focuses on a range of evaluation approaches from the methods, use, values, and social justice branches of evaluation. This will provide students with the opportunity to discuss a diverse range of paradigms and ways of constructing knowledge, range of methods, values stances, evaluation use (and misuse), evaluation theorists, and methodological assumptions. Our goal in this course is to challenge students to think about the role of evaluation as promoting social betterment (Henry, 2000; Mark, Henry, & Julnes, 2000) by thinking through the implications of their work as evaluators and researchers, and taking the time to explore the many (and varied) consequences of evaluation (and the role of evaluation) in society.

CONCLUSION
In this practice note, we shared the findings of a recent study we did with our students as they transitioned from the world of the classroom to active engagement with the dynamics of evaluation practice. Our goal in this study was to better understand how students navigate between the world of theory and the real world context of evaluation as a way to inform our teaching practice. As new evaluation teachers, we were interested in exploring the pedagogic implications of our research with students so that we could ultimately provide a more dynamic and focused learning experience for our students. Our research led to the subsequent redesign of our course offerings to provide a more nuanced transition between two dichotomized conceptualizations of evaluation—theory and practice. As we continue to reflect on our teaching and on our students’ experiences inside and outside the classroom, we expect to continue to revise our teaching practices to reflect our own experiences inside these dynamic pedagogical environments.

REFERENCES


doi: 10.3138/cjpe.31130


The Impact of Practice on Pedagogy


AUTHOR INFORMATION

Ayesha Boyce is an assistant professor within the Educational Research Methodology Department at the University of North Carolina at Greensboro (UNCG). Her research focuses on attending to value stances and issues related to diversity, equity, access, climate, cultural responsiveness, and restorative justice within evaluation—especially multisite, STEM, and contexts with historically marginalized populations. She also examines teaching, training, and novice learning in evaluation. Dr. Boyce has evaluated many programs funded by the National Science Foundation, the National Institutes of Health, Arizona Department of Education, Title VI, Illinois P-20 Council, and Spencer and Teagle foundations. She currently sits on the American Evaluation Association (AEA) Thought Leaders Committee for “An AEA Dialogue on Race and Class.” From 2014 to 2016 she was the chair of the AEA STEM Education and Training TIG. She teaches classes in program evaluation and research methodology where she emphasizes good practice with mindful attentiveness to theoretical roots.

Jill Anne Chouinard is an assistant professor within the Educational Research Methodology Department at the University of North Carolina at Greensboro (UNCG). Her main interests are in cross-cultural/culturally responsive approaches to research and evaluation, participatory research and evaluation, and evaluation and public policy. Dr. Chouinard has extensive experience working on evaluations at the community level in the areas of education and training, social services, health, and organizational learning and change. Much of her evaluation work has been conducted in First Nations and Inuit communities, as well as in other culturally diverse community settings. She maintains a strong connection between her evaluation practice and her academic research. As such, her research and academic writing focuses on participatory approaches to evaluation in culturally diverse community settings, from both a national and an international perspective.
Chaplowe and Cousins’ book, Monitoring and Evaluation Training: A Systematic Approach, is a timely and useful addition to the growing field of evaluation capacity development. Building monitoring and evaluation capacity—creating evaluation units, identifying M&E roles and responsibilities, and developing key stakeholders’ understanding of the use of M&E—is a challenge facing most countries today, particularly in light of the new Sustainable Development Goals (SDGs).

To respond to this challenge, there has been a significant rise in the quantity and quality of formal and informal national and international training programs (including masters and Ph.D. level programs). Discussions about the accreditation of individual evaluators, the establishment of International Development Evaluation Associations (IDEAs), national evaluation associations, and networks of evaluation associations are all important steps in building national and global monitoring and evaluation capacity. At the centre of much of this work are educators and trainers who can provide the knowledge and skills that would support the development of both the suppliers of monitoring and evaluation as well as those who need to be able to use evaluations.

This manual’s contribution is systematically exploring and providing guidance on how to use the best education and learning practices to facilitate the learning of monitoring and evaluation theory and practice.

Overall, the book is a useful tool for trainers who seek to prepare and deliver M&E courses and a solid resource for M&E practitioners to draw upon in the delivery of technical assistance. The authors pursue a system approach for training that makes a difference, importantly pointing out that training does not happen in isolation but as part of a larger system composed of interdependent parts. As such, to understand any one part of the system, it is best to examine it in relation to the overall system. M&E training is thus a system within a system and is dynamic and changing.

The authors argue that delivering an effective training program requires orderly, careful planning to gather information, analyze training demand, needs, and resources, identify relevant objectives, and realistically evaluate these objectives.

**Corresponding authors:** Charles Lusthaus, clusthaus@universalia.com; Hussein Amery, hamery@universalia.com.
Part 1, “Key Concepts of M&E Training,” provides a road map toward an understanding of the key concepts for M&E training. It reminds us that monitoring, evaluation, and the skills associated with evaluative thinking are acquired competencies and as such can be understood within institutional and organizational contexts as well as the best knowledge associated with adult learning theory. This section does both and gets you thinking.

Importantly, the authors consider the larger program management system for M&E training from the perspective of Results-Based Management (RBM). They correctly point out that current M&E evolved from RBM in the 1990s, coinciding with the demands for performance accountability and the adoption of the logical framework approach to programming by the public sector and international development agencies. “By identifying in advance the intended results of a program and how to measure them, M&E becomes more straightforward, and related reporting becomes more relevant and useful” (p. 11). Useful figures on the Basic Project Cycle and the Project Cycle with examples of M&E activities make this concept understandable and help the reader appreciate how the system works and where M&E “fits.” This section is accompanied by cartoons and numerous terminology tips, to make sure the reader has clarity in appreciating differences in approaches, types, and contents.

Selecting the proper training approach is often the key to success. Guidance is provided in the section “What Kind of M&E Training?”, which details the differences between short- and long-term training, face-to-face, organizational v. nonorganizational, and inclusive M&E training. Alongside this are useful suggestions on the scope of training (Basic Volkswagen v. Luxury Mercedes Benz) and the design of training to meet specific types of learners. Table 2.2, “Comparing the Three Major Training Categories,” provides the readers with a clear vision of the differences between distance delivery (asynchronous and synchronous) and face-to-face delivery, enabling an understanding of which option is most suitable for selection.

Sections on M&E capacity building and adult learning further support understanding the importance of the differences between individual and organizational levels of capacity building and external environment capacity building. While the authors underpin this by pointing out the importance of capacity assessment and strategic planning, they indicate it is beyond the scope of the book; thus, only limited attention is given to this aspect. Extensive and well-detailed attention is, however, provided on how adult learning takes place, learner-centred training, and principles for effective M&E training. For those without experience and knowledge in andragogy, this section will be useful in developing and transferring M&E knowledge to adults. For experienced adult trainers, the section will provide a solid overview.

Part 2, “A Systematic Approach to M&E Training,” provides potential trainers with a systematic approach to providing training. After introducing a variety of frameworks, the authors present the ADDIE training framework and use this to explain in great detail the steps required to develop, implement, and evaluate
monitoring and evaluation training programs. This section is full of examples, activities, checklists, and interesting textboxes designed to support trainers’ design and delivery efforts.

Starting with Chapter 6, “An Overview of the Addie Training Framework,” Section 2 represents the depth of the book and its systems approach. Chapter 7 walks trainers through various aspects of training analysis that will enable the next step. Chapter 8, “Training Design.” Following the authors’ approach, design starts with identification and preparation of training and learning objectives, followed by organizing instructional content and designing training curriculum, culminating with designing training evaluation and follow-up. The strength of the book is once again exhibited with excellent and comprehensive models upon which to build, in this case an easy-to-understand example curriculum template for training planning and example topics for M&E training curriculum.

Chapter 9, “Training Development and Preparation,” takes the reader to the workshop stage of the training process. Developing sound and usable materials are essential for uptake and application by learners. The authors emphasize that the key to ensuring the trainer is ready to deliver by noting the importance of material review and piloting to ensure preparedness. Examples of materials and resources for M&E training and outlines for training strategy, syllabus, and lesson plans are provided; importantly, the authors note that in each case these must be tailored to the needs and contexts of learners. This is nicely rounded out by a detailed training preparation checklist that supports facilitators in ensuring they are ready to deliver. Chapter 10, “Training Implementation,” is well detailed and focused on delivery of training; however, this area of the book could apply to training in any context and not specifically M&E training.

Chapter 11, “Training Evaluation,” focuses on the last stage of the ADDIE model for training. It is designed to ensure that training objectives are aligned with measurable results. The authors define training evaluation as a “systematic investigation of the worth or merit of training.” They rightly point out the similarities to training debrief and, citing Michael Patton, they indicate the same three questions are being asked for evaluation: what, so what, and now what. This is the essence of their approach to evaluation and it offers a sequence of inquiry of what happens during training, what difference it makes, and what recommendations can be made for future training. A useful overview of training evaluation methods is provided, with particular emphasis on the “levels approach” to training evaluation, including examples from Kirkpatrick’s 4 Levels and Guskey’s Five Levels for Evaluating Professional Development. The authors provide an interesting example of a training logframe, which is a sound approach to developing results-based training. The appendices in the section contain substantial and useful tools for evaluation data collection methods, M&E learning development plans, training feedback forms, and training assessment forms.

Part 3, “M&E Training Methods and Techniques,” provides the reader—and particularly M&E trainers—with a wide variety of exercises, facilitation and teaching techniques, learning methods, and assessment techniques. This part of the
book is a tool box with 99 real-life examples. These M&E training methods and techniques provide clear, practical, and useful advice on planning and delivering training. Sixteen different learning activity approaches are provided, alongside review assessment and M&E activities. The authors achieve their objective in this part of the book, which is a selection of activities to inspire active, engaging delivery of M&E. The learning games section introduces fun and innovative activities such as “M&E Masterpiece,” the “SMARTer” Indicator game, “M&E Puzzle Games,” and “M&E Jeopardy,” which are innovative and differ from the usual approaches. These are introduced alongside other more traditional approaches such as practicum, guest presenters, and independent learning activities. The strength of the section is the brevity, clear presentation, and ease of understanding, so anyone of the approaches could be easily applied after reading the book. Photos and visuals aid in understanding how such approaches might be put into practice.

The section on training closing activities reminds the reader of the importance of bringing proper closure to training and helping trainees organize and remember training content. The section is more of a menu to select from than an innovative introduction. As the authors point out, closing activities done poorly or in a rushed manner leave trainees with a negative impression of the experience. A variety of options are thus provided that can enable the trainer to select an approach that is suitable for the context. Fun approaches such as role-playing are mixed with more standard methods of Senior Management Debrief, the Evaluator Speech, “emptying the parking lot” (where pending items are dealt with), Affirmation, Mingle or Web, or the standard awards ceremony.

Training follow-up activities provides helpful guidance to demonstrate that delivering training doesn’t end at the workshop. Fourteen subareas ranging from suggestions for the use of social media platforms, peer-reviewed learning assignments or work samples, distance and e-learning follow-up, and trainee development plans to packaging M&E training as part of larger M&E capacity-building exercises offer the reader a useful roadmap to ensuring the completeness of service delivery to clients.

The lengthy references provided in the book will aid M&E professionals looking for additional reading to support their work in the field.

In general, this book is a good resource and reference for both organizational professional development and classroom education. As evaluation monitors and consultants, we have put together for our practice a wide assortment of material helpful for our training. However, we have been waiting for a book with this broad, in-depth coverage. Every aspect of the work of the trainer-evaluator is covered, and it is clear that both authors are personally experienced in this field. Overall, the book is very easy to use, with succinct colour-coded chapter summaries and recommended resources at the end of each chapter, which makes the book a useful one to anchor training delivery.

As in many such books, decisions needed to be made with respect to what should be included and left out. I was somewhat disappointed that more attention was not paid in Part 1 to the interwoven nature of monitoring and evalua-
tion, and similarly, the role planning and use play in the success of the evaluation enterprise. However, these can be the subjects of the next book written by these highly skilled authors.

As the authors note, the book was written to address potential obstacles and resistance to M&E training, a resistance that stems from a lack of understanding of the subject and the value of M&E. As such, the book is designed so that M&E training can be enjoyable, meaningful, and engaging for learners, so that it can be “understood, appreciated and used.” Chaplowe and Cousins have achieved this and delivered a comprehensive and easy-to-use learning tool. The numerous well-detailed figures, sample activities, tables of lesson plans, and examples of tools provider trainers with a plethora of options in preparing and delivering comprehensive M&E training.

NOTE

1 Canada is the first country to institute a national certification provided by its professional association.
A fundamental issue facing evaluators, particularly new evaluators, is how to attend to quality within one’s craft. Through coursework and on-the-fly training, new evaluators strive for proficiency in current evaluative practice, but may fail to position this learning within a bigger picture of the “how” and “why” of what evaluators do. This book invites readers to consider deeply and critically the many factors impacting evidence quality—including how one can promote accurate and meaningful evaluation use.

OVERVIEW

As a whole, the book is a call to expand the contemporary concept of evidence in evaluation beyond one determined solely by method choice. The authors demonstrate the theoretical need for such an expanded focus and discuss how various methods impact evidence credibility. Taken together, the chapters facilitate a renewed examination of the concept of evaluation evidence and lay the groundwork for translating this call into action.

Part I provides a primer on the debate surrounding the definition and operationalization of quality in evaluation. The introduction addresses the history of the debate (Chapter 1, Donaldson), the nature of quality relative to the philosophy of science (Chapter 2, Christie/Fleischer), and “peripheral factors” impacting judgments of credibility (Chapter 3, Miller). The theoretical foundation offered in Part I positions evidence quality relative to context, research questions, stakeholder values, and guidelines for evaluation (utility, feasibility, accuracy, and propriety). Because credibility must be established with respect to the impact (p. 28) and influence (p. 41) of an evaluation, “peripheral factors” above and beyond “the facts” influence credibility judgements. Thus, evaluators can provide rigorous and influential evaluations (p. 23) by considering critically the relevant factors impacting evidence generation and use.

Part II focuses on experimental designs, specifically the strengths and limitations of the randomized controlled trial (RCT) design. An illustration

Corresponding author: Robyn Thomas Pitts, pitts.robyn@gmail.com

© 2017 Canadian Journal of Program Evaluation / La Revue canadienne d’évaluation de programme
of the political backstory surrounding RCTs (Chapter 4, Henry) is followed by a discussion of susceptibilities, practical issues, credibility, and causality related to RCTs (Chapter 5, Bickman/Reich). The section concludes with an explanation of causal concepts and underlying logic, myths, and alternative designs (Chapter 6, Scriven). Some in the field consider RCTs the “gold standard”—the most credible design for establishing causality. While RCTs are tremendously feasible research designs (Chapter 4), they are susceptible to numerous threats (i.e., generalizability, practical considerations, ethical concerns). Besides, not all research questions can be informed by experimental designs. Therefore, judging research credibility cannot be without consideration of contextual information: the purpose of the study, the research questions, and details regarding what evidence is generated, how it is generated, and for whom it is generated (Chapter 5).

In Part III the focus shifts to the use of qualitative methods for producing credible and actionable evidence (Chapter 7, Rallis), delving into how nonwritten artifacts (i.e., images) may serve as credible evidence (Chapter 8, Mathison). Finally, evaluation synthesis is described as a methodology for answering evaluation questions with the intent of maximizing the perceived credibility of findings (Chapter 9, Chelimsky). After scrutinizing credibility in experimental designs, Part III argues that the potential for evidence to be credible and actionable concerns more than methodology. Evidence quality results from the ethical and rigorous application of methods in alignment with the purpose of the evaluation in such a way that the reasoning behind method selection is transparent (pp. 137–138). Analysis of the use of images as evidence demonstrates that “any evidence may be credible, truthful, and useful” (p. 174), given sufficient alignment between the evidence and the evaluation questions and criteria. Overall, Part III makes the case that methodological credibility is merely one facet of credible evaluative evidence (p. 181) and that evaluators should be mindful of how additional factors, such as evidence presentation (p. 184), can influence perceived credibility.

Part IV broadens the discussion by highlighting nonmethodological factors impacting evidence, followed by a discussion of context-dependent relationships and interactions affecting credibility (Chapter 10, Greene), five dimensions of context (Chapter 11, Julnes/Rog), and the importance of a practical orientation (Chapter 12, Schwandt). The final chapter (Chapter 13, Mark) offers a thematic synthesis and framework for attending to evidence quality. While Parts II and III focus on evidence credibility, Part IV elaborates evidence actionability and provides practical guidance for evaluators. Actionability comprises four criteria (credibility, inferential potency, relevance, and comprehensiveness of evidence) and should impact method selection (Figure 13.1, p. 279). Many aspects of context and tiers of processes influence these four criteria (Figure 13.2, p. 288), establishing a “threshold of actionability” for deciding to act in response to evidence. The text concludes by offering a 7-step process for addressing evidence quality and suggesting further research studies.
RELEVANCE
The authors make a case for advancing the credibility dialogue beyond discussion of methods to a more nuanced interrogation of factors impacting evidence quality. The overarching idea is that rigorous methods used to generate information are necessary but insufficient to ensure quality—other factors matter. Further, because evaluations inform real-world situations, deference to methodological purity and a detached researcher mentality is not possible. Instead, the evaluator must recognize that “all evaluation is interested evaluation, serving some interests but not others” (Greene, 2006, p.135). To create high-quality, useful information, evaluators must navigate responsively rather than follow plans rigidly, must recognize values rather than adhere to technical processes, and must negotiate options rather than dictate the meaning of results.

The discussion of experimental and nonexperimental approaches highlights the specific role of evaluators in addressing context, stakeholder plurality, and diverse information needs throughout the entirety of an evaluation. Though evaluators are not necessarily experts in the content of the programs they serve, they are regarded as experts in the use of methodologies for doing evaluation. Thus the evaluator has an obligation to educate their client apropos which modes of inquiry best facilitate gathering and examining evidence, given the context and aims of the evaluation. In assuming this responsibility, evaluators shoulder the additional obligation of staying abreast of contemporary developments and discussions in the field of evaluation methodology. By giving a thorough treatment of contemporary views around the production of evaluation evidence by leaders in the field, this volume presents practicing evaluators a strong defense against the normative pressure to use a methodology not because of its ability to generate useful findings in the context at hand, but because of opinions and assumptions about the general value or rigour of certain methods over others. The book presents a strong case for method selection based on the needs and capacities of the evaluation context, as opposed to method selection based on decontextualized assumptions about the general strength or utility of one method compared to another.

CONTRIBUTION
Overall, this book provides a valuable resource for various audiences. For students, the text offers an opportunity to contemplate the purposes and implications of evaluation, the myths and misunderstandings within the field, and the role of evaluators in contemporary society. Instructors will find this text a useful tool for introducing students to the field. Readings will acquaint students with key issues in evaluation and will engage students in critical analysis of ontological, epistemological, and practical factors impacting evidence quality throughout evaluation design, implementation, and use. The text offers practitioners an analysis of credibility in light of various considerations and sound guidance on how to select those methods that yield credible and actionable evidence in context. Furthermore, it prompts practitioners to engage in careful thinking about evaluation assumptions.
and encourages readers to embrace a broader, more holistic consideration of the concept of evidence in evaluation.

REFERENCE