



paramètres

Sous la direction de
Valéry Ridde et Christian Dagenais

Approches et pratiques en évaluation de programmes

NOUVELLE ÉDITION REVUE ET AUGMENTÉE



Les Presses de l'Université de Montréal

**APPROCHES ET PRATIQUES
EN ÉVALUATION DE PROGRAMMES**

Nouvelle édition revue et augmentée

Sous la direction de Valéry Ridde
et Christian Dagenais

**APPROCHES ET PRATIQUES
EN ÉVALUATION DE PROGRAMMES**

Nouvelle édition revue et augmentée

Les Presses de l'Université de Montréal

Catalogage avant publication de Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada

Vedette principale au titre:

Approches et pratiques en évaluation de programmes

2^e éd.

(Paramètres)

Comprend des réf. bibliogr.

ISBN 978-2-7606-2782-6

1. Évaluation de programmes. 2. Recherche évaluative (Programmes d'action sociale).
3. Évaluation de programmes - Cas, Études de. I. Ridde, Valéry. II. Dagenais, Christian.
- III. Collection: Paramètres.

H62.A66 2012

361.2072

C2012-941099-3

Dépôt légal : 3^e trimestre 2012

Bibliothèque et Archives nationales du Québec

© Les Presses de l'Université de Montréal, 2012

ISBN (papier) : 978-2-7606-2782-6

ISBN (epub) : 978-2-7606-3125-0

ISBN (epub) : 978-2-7606-3124-3

Les Presses de l'Université de Montréal reconnaissent l'aide financière du gouvernement du Canada par l'entremise du Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour leurs activités d'édition.

Les Presses de l'Université de Montréal remercient de leur soutien financier le Conseil des Arts du Canada et la Société de développement des entreprises culturelles du Québec (SODEC).

IMPRIMÉ AU CANADA EN AOÛT 2012

Présentation

L'idée d'offrir en français un ouvrage pédagogique de référence sur l'évaluation de programmes était en germe depuis longtemps. Le succès qu'a connu l'ouvrage dès sa publication en 2009 a confirmé notre intuition au sujet de sa pertinence, car celui-ci était déjà en réimpression moins d'un an après son apparition sur le marché. Cela nous a amenés à préparer cette deuxième édition revue et enrichie de plusieurs chapitres.

Il existe encore un très petit nombre d'ouvrages avec des visées pédagogiques adaptées aux étudiants et aux professionnels qui participent à des formations en évaluation de programmes. S'il existe dans nos bibliothèques de nombreux livres francophones consacrés à ce domaine, tous sont issus de disciplines particulières telles que les politiques publiques, les services sociaux, l'administration publique, la santé ou encore le développement international. Les textes courts sont certes assez nombreux, mais épars et difficiles à trouver. Bien souvent, ils sont noyés au sein d'un ouvrage traitant d'un sujet plus général ou publiés dans une revue non spécialisée. De fait, une seule revue scientifique consacrée à l'évaluation publie des articles en français, la *Revue canadienne d'évaluation de programmes* ; mais au moment où nous rédigeons cette deuxième édition, on entend parler de l'arrivée d'une revue africaine d'évaluation qui, si le projet se concrétise, saura, nous l'espérons, laisser une large place au français.

Régulièrement placés devant ces lacunes dans notre enseignement universitaire et lors de formations professionnelles menées au Québec, en

France, en Haïti ou en Afrique, nous avons conçu le projet de publier le présent ouvrage. Sa planification et sa réalisation se sont étalées sur près de deux ans pour la première édition (2009), le temps de trouver les thèmes principaux et, surtout, les auteurs clés, des spécialistes disposés à se plier aux exigences didactiques que nous avions élaborées. Tous les chapitres de ce livre ont donc été écrits sous une forme accessible par des pédagogues, des enseignants universitaires et des formateurs rompus depuis de longues années à l'exercice du partage de connaissances en évaluation de programmes tout en mettant l'accent sur la pratique plutôt que sur la théorie. Nous avons même conclu une entente en 2010 avec les Presses de l'Université de Montréal et L'Harmattan Burkina Faso pour rendre financièrement accessible l'édition africaine vendue uniquement sur ce continent. Aujourd'hui, trois ans après la publication du livre, les nombreux commentaires apportés par les lecteurs et les collègues nous ont incités à produire cette deuxième édition. Nous avons notamment ajouté quatre nouveaux chapitres théoriques, car les connaissances en évaluation évoluent constamment et nous voulions que les lecteurs de langue française puissent avoir accès aux dernières acquisitions dans le domaine. Il manquait également dans la première édition des exemples relatifs à l'usage des méthodes mixtes, pourtant décrites dans la première partie. Deux nouveaux chapitres viennent donc combler cette lacune. Ils se combinent avec le chapitre théorique sur les méthodes mixtes, qui a été entièrement mis à jour.

Dans la première partie, le lecteur approfondira sa compréhension de certains concepts dont il aura pu appréhender sommairement les contours dans une introduction générale. L'objectif est de montrer que le domaine de l'évaluation de programmes fait appel à un nombre important de théories, d'approches et d'outils pour répondre aux questions posées par les commanditaires. En effet, l'évaluation puise dans différentes disciplines telles que l'étude des politiques publiques, l'épistémologie ou encore la statistique. Dans cette nouvelle édition, les lecteurs pourront se familiariser avec la stratégie de l'étude de cas, l'évaluation économique, les approches participatives et l'approche *Realist*.

Un défi essentiel auquel fait face tout enseignant en évaluation est lié à la maîtrise de la grande diversité des approches évaluatives et des types d'évaluation. La seconde partie de l'ouvrage présente quelques études de cas choisies pour montrer clairement comment les concepts qui auront

été exposés sont employés dans la pratique. Ces chapitres recouvrent plusieurs domaines disciplinaires et proposent divers exemples de pratiques évaluatives. Chaque étude suit un plan similaire permettant de rendre la description des cas relativement homogène. Le tableau 1.1 qui figure dans l'introduction permettra au formateur de repérer, selon les types d'évaluation, les différentes études de cas correspondantes.

Les cas présentés sont tirés de contextes socioéconomiques variés, ce qui permettra à ceux et à celles qui liront ce livre au Canada, au Mali ou en France, par exemple, d'appréhender des réalités différentes. Dans cette nouvelle édition sont introduits deux cas illustrant l'utilisation des méthodes mixtes au Québec et au Burkina Faso.

Nous souhaitons remercier les auteurs, qui ont répondu favorablement à nos requêtes et ont su se placer à la portée des étudiants pour partager leurs connaissances, y compris dans la mise à jour de cette nouvelle édition. Sans leurs efforts, ce manuel n'aurait évidemment pu voir le jour.

Valéry Ridde et Christian Dagenais

PREMIÈRE PARTIE

CONCEPTS ET APPROCHES

1

Introduction générale à l'évaluation de programmes

Valéry Ridde et Christian Dagenais

L'évaluation n'est pas chose nouvelle. On raconte dans l'Ancien Testament qu'à la cour du roi de Babylone, Nabuchodonosor, le chef des eunuques organisa une évaluation que l'on qualifiera ici de quasi expérimentale (voir chapitres 6 et 9). Il s'agissait d'étudier les effets différenciés d'un régime alimentaire sur la santé des enfants hébreux de la cour. Mais au-delà de la légende, il faut reconnaître que le champ de l'évaluation de programmes s'est surtout développé dans les trente dernières années. L'origine de la pratique et de la recherche en la matière se situe essentiellement dans le domaine de l'éducation. Les enseignants se sont interrogés sur l'influence que pouvaient avoir leurs cours sur les connaissances acquises par leurs élèves. En changeant d'objet, passant des habiletés et des connaissances acquises par les individus aux programmes qui organisaient cette transmission, le champ de l'évaluation s'est transformé pour devenir beaucoup plus interdisciplinaire, voire transdisciplinaire pour certains. Ainsi, quand on cherche à appréhender pour la première fois le domaine de l'évaluation de programmes, on s'aperçoit qu'il existe de multiples façons de le concevoir. Si vous consultez cinq ouvrages de référence, vous risquez d'y trouver cinq manières distinctes de définir l'évaluation et ses concepts. Il s'agit en effet d'une discipline dont les concepts et la pratique n'ont pas encore été clairement codifiés. Voilà pourquoi

plusieurs sociétés nationales d'évaluation cherchent encore à se doter de chartes, de codes de déontologie, de normes de pratique ou de compétences essentielles, alors que d'autres en ont adopté récemment (voir chapitre 12). Voilà aussi pourquoi, lorsque l'on souhaite renforcer les capacités et les pratiques évaluatives au sein d'une organisation, il est toujours bon de se doter d'une politique commune d'évaluation, voire d'un glossaire des termes que l'on utilisera (le présent ouvrage en propose un en page 433).

La présente introduction est fondée sur notre propre expérience de l'enseignement et sur notre pratique de l'évaluation dans un grand nombre de contextes et sur diverses problématiques. Il ne s'agit évidemment pas de suggérer une nomenclature définitive ou supérieure aux autres. Notre objectif est bien plus modeste et consiste à présenter au lecteur le champ de l'évaluation de programmes afin de faciliter sa lecture des différents chapitres de cet ouvrage et éventuellement le diriger vers les références essentielles dans le domaine.

Les courants de pensée : une brève histoire de l'évaluation

Les courants de pensée en évaluation sont nombreux, et la façon dont les évaluateurs se représentent le monde (leur « paradigme ») conditionne largement les questions qu'ils se posent et les stratégies qu'ils emploient pour y répondre (Bégin *et al.*, 1999). Nos valeurs, croyances et intérêts interfèrent assurément dans nos pratiques professionnelles. Dans le cas qui nous occupe, ils déterminent les choix théoriques et méthodologiques de la pratique de l'évaluation. Il ne s'agit pas de discuter longuement des grands paradigmes épistémologiques, mais il est essentiel d'en dire quelques mots pour que le lecteur, évaluateur actuel ou en devenir, soit en mesure de s'interroger à son tour sur ses propres choix¹. Pour le définir simplement, un paradigme, selon la définition classique de Thomas Kuhn, est un système de croyances partagé par une majorité d'individus œuvrant au sein d'une même discipline². Ce système de référence est un cadre

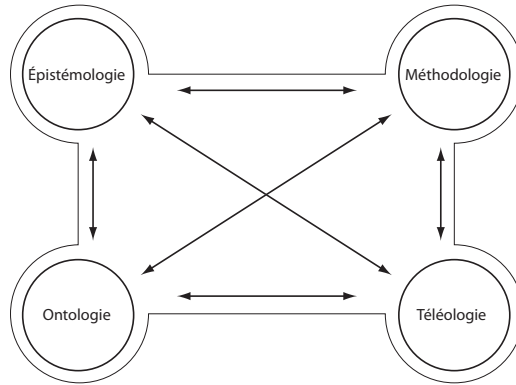
1. Pour approfondir cette réflexion, nous suggérons la lecture de l'excellent manuel d'introduction à l'épistémologie de Fourez (2003).

2. Hall (1993) a aussi employé ce terme dans son concept de paradigme politique, certainement utile aux évaluateurs des politiques publiques (voir chapitre 4 et l'ouvrage d'analyse politique récemment publié par Paquin *et al.*, 2011).

normatif qui oriente les questions, les méthodes (voir chapitre 7) et la manière d'attribuer un sens aux données que l'on recueille. À la suite de Levy, Gendron (2001) propose qu'un paradigme soit constitué de quatre champs interreliés.

FIGURE 1.1

Les quatre champs constitutifs d'un paradigme



Source : Gendron, 2001.

- Le champ épistémologique renvoie à la vision du monde de l'évaluateur, la légitimité des analyses qu'il produit et surtout sa relation (et sa distance) avec les données qu'il recueille et analyse (subjectivité *par opposition* à objectivité).
- Le champ ontologique englobe la nature du monde vécu et la compréhension qu'on a de la manipulation possible par l'évaluateur des entités physiques ou mentales. C'est de la construction de la réalité dont il est ici question (une seule réalité *par opposition* à plusieurs).
- Le champ méthodologique rassemble les techniques que l'évaluateur emploie pour appréhender le monde vécu.
- Le champ téléologique regroupe les questions relatives aux finalités des évaluations ainsi qu'aux intérêts et intentions des évaluateurs d'un programme.

Le paradigme au sein duquel l'évaluateur s'inscrit conditionne les méthodes utilisées, par exemple, pour répondre aux questions d'évaluation posées par les responsables d'un programme. Certains évaluateurs, dits positivistes, ne pourront croire à l'efficacité d'un programme que dans

la mesure où il aura été possible d'en comparer les effets, comme dans un laboratoire, entre un groupe de participants (cas) et un groupe de non-participants (témoins), l'appartenance de ces personnes à ces groupes ayant été déterminée aléatoirement (voir le chapitre 6). D'autres évaluateurs, plutôt constructivistes, affirmeront à l'inverse qu'il est impossible d'isoler les effets d'un programme de son environnement, et qu'il faut employer une méthodologie permettant de relever ces interactions. Ils considèrent ainsi que dans le cas des programmes sociaux, il est impossible de reproduire des conditions de laboratoire. Ils soutiennent aussi que toute connaissance sur les effets d'un programme est construite conjointement par les acteurs sociaux concernés. Ces évaluateurs constructivistes assument pleinement la subjectivité de leur jugement, puisqu'ils estiment que l'objectivité dont se réclame le camp positiviste n'est pas atteignable. Les tenants de l'approche réaliste en évaluation (Blaise *et al.*, 2010), optant pour une perspective épistémologique intermédiaire, postulent de leur côté que la réalité existe en dehors des constructions humaines, mais qu'elle n'est perceptible qu'à travers nos sens (voir chapitre 14).

Les différents courants de pensée de l'évaluation se sont développés tout au long de son histoire. L'une des manières de décrire cette histoire, car nous sommes d'accord avec Shaw et ses collègues (2006) pour dire qu'il en existe plusieurs, est de suggérer l'existence, parfois concomitante, de quatre générations d'évaluation (Guba et Lincoln, 1987). La première génération est celle de la mesure, dont l'évaluateur est un spécialiste. La deuxième est plutôt de l'ordre de la description. L'évaluateur s'attache surtout aux processus qui permettent la survenue des effets constatés. Dans la troisième génération, le défi de l'évaluateur est de porter un jugement sur les résultats obtenus en regard des objectifs initialement fixés. Enfin, dans les années 1970 est apparue une quatrième génération, qui se veut bien plus interactive. Les évaluateurs usent de la négociation, deviennent des facilitateurs entre les instances impliquées. Il s'agit de reconnaître non seulement le point de vue et les valeurs de l'ensemble des acteurs concernés, mais aussi de faire en sorte qu'ils fassent partie intégrante de ce processus. Autrement dit, il s'agit d'évaluations pluralistes dont l'approche participative, décrite plus bas, est l'une des stratégies possibles (voir chapitre 13).

L'évaluation de programmes : une définition

Plutôt que de commencer par une définition universitaire, nous vous proposons un bref séjour au soleil, dans un pays de l'Afrique de l'Ouest. Par une radieuse journée de janvier, un expert québécois anime une séance de formation à l'évaluation de programmes au Burkina Faso. Cette séance est organisée à l'intention des sages-femmes et des intervenantes d'une organisation non gouvernementale (ONG) qui met en œuvre un programme de maternité dite « à moindre risque ». En commençant sa formation sur l'évaluation de programmes, l'expert interroge les participantes sur les mots qu'elles associent au terme « évaluation ». On évoque le contrôle, l'audit, la mesure, la vérification, etc. Des Canadiens présents dans la salle parlent de l'idée de reddition de comptes. La teneur de ce lexique ne surprend guère le formateur, puisque aujourd'hui encore, beaucoup associent avant tout l'évaluation à des fonctions de contrôle qui attribuent de bonnes ou de mauvaises notes à un programme... et à ses responsables. Puis, l'expert demande : « quelle température fait-il aujourd'hui ? » Les agentes de la santé trouvent qu'il fait plutôt froid, les expatriées canadiennes qu'il fait chaud, et l'expert québécois qui vient d'arriver juge qu'il fait très, très chaud. « Comment avez-vous fait pour juger de la température et dire qu'il fait froid ou chaud ? ». Chacun répond de manière différente à la question de l'expert : « c'est l'hiver », « j'ai regardé le thermomètre », « j'ai comparé avec la température de la veille », « je transpire beaucoup plus aujourd'hui qu'hier », « c'est la saison fraîche ». Puis, les participantes admettent toutes avoir choisi leurs vêtements en fonction de la température de ce matin-là. Les personnes présentes à cet atelier de formation ont alors pris conscience du fait que l'évaluation est un acte permettant de porter une appréciation (froid, chaud, tempéré) selon une démarche critique fondée sur une collecte systématique de données (température, sudation, saison) à propos de multiples objets (la température) dans le but de prendre des décisions (comment s'habiller).

Certains auteurs soutiennent que l'appréciation évaluative à l'égard d'un programme doit être posée selon une démarche scientifique ou en ayant recours à des méthodes scientifiques. Nous préférons ne pas employer le terme « scientifique », car dans notre pratique de l'évaluation, le mot a toujours eu tendance à créer une certaine distance avec les parties impliquées dans l'évaluation. L'évaluateur est alors rapidement perçu comme un « chercheur », avec toutes les connotations qui, à tort ou à

raison, s'y rattachent. En outre, évoquer la science fait perdurer la difficulté qu'ont la plupart des auteurs à différencier la recherche de l'évaluation (chapitre 2). La différence fondamentale, selon nous, entre la recherche et l'évaluation est que la finalité *principale* de la première est de produire des connaissances, alors que le but de la seconde est de fournir une information utile sur le fonctionnement d'un programme ou pour permettre une prise de décision le concernant.

Évidemment, les choses ne sont pas toujours aussi tranchées : on peut parfois produire des connaissances à l'aide d'une évaluation, ou prendre des décisions à la suite de travaux de recherche. Mais il reste que le choix de ne pas la définir explicitement comme une science ne signifie pas que l'évaluation soit une démarche frivole où l'appréciation que l'on porte sur l'objet programme n'est pas fondée. Au contraire, la rigueur et la systématisation en sont les deux valeurs-phares.

L'analogie climatique est simple, mais suffisamment claire pour définir l'acte d'évaluation et s'assurer que les personnes avec qui nous allons travailler ne la perçoivent pas comme un processus de contrôle (bien que cela puisse parfois être le cas). Évaluer la température est un peu plus simple que d'évaluer un programme, surtout lorsqu'il est complexe comme le sont la plupart des interventions sociales (voir chapitre 14). Mais comment peut-on définir un programme ? Ici encore, une multitude de définitions se croisent. Nous retenons celle de Plante (1994). Bien qu'un peu longue, elle a le mérite de préciser clairement les contours de ce qui constitue un programme :

[Un programme est un] ensemble cohérent, organisé et structuré d'objectifs, de moyens et de personnes qui l'animent. Il se justifie sur la base de besoins définis comme une carence ou un manque qui affecte des individus, une collectivité ou une société. Il est sous le contrôle d'une ou de plusieurs personnes responsables de la qualité de sa formulation et de son fonctionnement. Il est mis en place pour transformer des choses ou l'état d'une chose.

Le maître mot de cette définition est *transformation*. Un programme est donc un ensemble organisé et cohérent d'activités dont l'objectif n'est pas d'exister pour lui-même, mais plutôt de produire des changements pour les participants potentiels. En effet, il ne sert à rien, ou alors à d'autres choses que celles qui sont évoquées ici, de développer un programme de maternité à moindre risque si celui-ci vise seulement à construire des maternités, à former des sages-femmes ou à rédiger des protocoles

thérapeutiques. Ce que l'on vise avant tout, c'est l'obtention à court terme d'une meilleure qualité des soins pour les parturientes et leurs nouveau-nés, et à plus long terme, une réduction de la mortalité maternelle et infantile (voir chapitre 18). Il faut aussi comprendre que tout programme peut produire des changements inattendus dont les évaluateurs devront aussi rendre compte (Morell, 2010). Ces effets peuvent avoir été anticipés ou non, et être bénéfiques, neutres ou nuisibles pour les participants.

Même si les frontières ne sont pas toujours faciles à tracer, cet ouvrage traite de l'évaluation de programmes, et laisse donc de côté l'évaluation des pratiques professionnelles, des connaissances individuelles ou encore des politiques publiques.

Au-delà des aspects conceptuels de ces définitions de ce qu'est un programme, il faut noter combien les acteurs sociaux jouent un rôle crucial dans sa dynamique. C'est par l'exercice du pouvoir que se réalisent ou non les programmes, bien loin de se déployer d'eux-mêmes (Crozier et Friedberg, 1977 ; Lemieux, 2002). Il faut donc être très attentif au jeu des acteurs, aux relations de pouvoir et aux interactions sociales. Aussi plusieurs acteurs sont-ils impliqués lorsqu'un programme est défini, puis mis en œuvre, mais également lorsqu'il est évalué. L'évaluation, en tant qu'activité humaine, est source de tensions et de relations de pouvoir. Il ne faudrait donc pas voir un programme et son évaluation comme une activité mécanique, uniquement technique, où ne nous intéressent que les objectifs et les indicateurs.

Si un programme peut aussi être conçu intellectuellement comme la résultante d'un ordre négocié entre des acteurs sociaux, selon l'expression d'Anselm Strauss, aux fins de cette introduction, il nous faut avoir recours à des subterfuges didactiques pour expliciter les concepts théoriques. Les différents termes que nous suggérons sont donc essentiellement heuristiques, car ils expliquent la matière enseignée. Il ne s'agit cependant pas de nier la complexité des programmes, dirait Edgar Morin, et des interactions sociales. Les tenants de l'évaluation réaliste discutée au chapitre 14 perçoivent le changement social comme un système de transformation, un système ouvert qui est le produit d'une myriade d'éléments et de forces impossibles à quantifier (Pawson, 2006 ; Ridde et Guichard, 2008 ; Blaise *et al.*, 2009). On aura donc compris que la vision du monde et de la science qu'ont les évaluateurs – leur perspective paradigmatique – conditionne leur manière de percevoir et d'évaluer un programme.

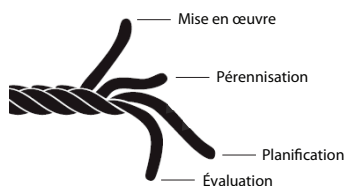
L'évaluation n'est pas l'étape finale d'un programme

La vision des programmes qui veut que ceux-ci doivent être adaptés vient contredire celle qui décrit leur vie selon un cycle inéluctable allant de l'identification des besoins à l'évaluation, en passant par les étapes de la planification et de la mise en œuvre. La plupart des manuels de gestion de programme illustrent encore le processus des programmes par un cercle représentant ces différentes étapes jointes par des flèches. Les fondements de cette conception remontent assurément aux « étapistes » des années 1950 qui étudiaient les politiques publiques. L'évaluation y est souvent l'étape ultime du processus. D'autres ajoutent parfois la pérennité ou encore la capitalisation (voir chapitre 10 pour un exemple) pour clore la boucle.

Nous estimons pour notre part que le processus des programmes n'est ni linéaire ni cyclique. À la suite des propositions théoriques énoncées par Pierre Pluye et ses collaborateurs (2004), il faudrait concevoir un programme comme étant composé d'une série de quatre sous-processus concomitants et interdépendants que sont la planification, la mise en œuvre, l'évaluation et la pérennisation. Ces quatre sous-processus pourraient ainsi être représentés par les quatre fils d'une même corde programmatique.

FIGURE 1.2

Concomitance et interdépendance des sous-processus d'un programme



Adapté de Pluye *et al.*, 2004.

La planification permet de prévoir l'allocation des ressources nécessaires pour atteindre des objectifs précis grâce à la mise en œuvre d'activités préalablement définies. Quant à l'évaluation d'un programme, adaptant des définitions plus classiques et formelles (Grembowski, 2001; Mathison, 2004; Patton, 1997) que celle que nous avons évoquée, nous

dirons qu'il s'agit d'un processus systématique de collecte d'informations utilisables pour apprécier un programme et ses composantes, dans le but de prendre des décisions. L'évaluation est le plus souvent une démarche d'amélioration des projets et des pratiques, en somme un moteur de changement (Ridde, Pommier et Jabot, 2010). Enfin, alors que la mise en œuvre d'un programme consiste en la mobilisation de ressources pour organiser des activités afin d'atteindre des objectifs, la pérennisation concerne le devenir des programmes. La pérennisation est en effet centrée sur le processus qui permet la continuation des activités et des effets des programmes. La pérennité, elle, est le résultat du processus de pérennisation – résultat qui se manifeste dans les routines organisationnelles – et peut être évaluée de manière transversale, à tout moment et ponctuellement, en fonction de degrés de pérennité. Les actions favorables à la pérennité doivent démarrer dès la mise en œuvre du programme, et non à la fin de celui-ci (Pluye *et al.*, 2005). Nous n'irons pas plus loin dans cette réflexion sur la pérennisation et la pérennité, certes nouvelle dans le domaine, mais dont les détails conceptuels et les illustrations empiriques sont présentés ailleurs (Pluye *et al.*, 2004; Ridde, Pluye et Queuille, 2006).

Retenons que ces quatre sous-processus sont concomitants. Ceux qui ont été impliqués dans des programmes savent bien que lorsque l'on met en place un programme (mise en œuvre), on s'interroge toujours sur ce qui se passe (évaluation). On revoit toujours ce qu'il a été prévu de faire (planification) en se préoccupant, ce qui demeure cependant souvent le maillon faible des programmes, de la façon de procéder pour assurer l'avenir de ces derniers (pérennisation).

La finalité d'une évaluation

Dans un livre publié au milieu des années 1980, Patton établit ce qu'il a nommé une « soupe à l'alphabet de l'évaluation » (*alphabet soup*), au sein de laquelle il finit par trouver 100 types d'évaluation de programmes. Plus récemment, il a réduit ce nombre à 58. Stufflebeam (2001) suggère quant à lui de regrouper 22 types d'évaluation en quatre grandes catégories :

- pseudo-évaluations : relations publiques ;
- quasi-évaluations : orientées vers les questions et les méthodes ;
- amélioration et imputabilité : orientées vers le mérite et la valeur ;
- plaidoyer social : orienté vers la différence dans la société.

Malgré ces distinctions variées, il est possible de suggérer des différences selon les finalités des évaluations. De manière classique, on estime que les évaluations poursuivent deux grands objectifs : rendre des comptes ou améliorer un programme. Le premier type d'objectifs permet aux responsables d'un programme de rendre des comptes dans une perspective d'imputabilité. On parle parfois d'une évaluation sommative, car elle est bien souvent réalisée à la fin d'un programme, soit *ex post*. Le second type d'objectifs produit des informations utiles pour corriger la manière dont les activités d'un programme sont mises en œuvre. On parle aussi d'évaluation formative, car celle-ci est habituellement effectuée pendant la mise en œuvre d'un programme. Au-delà de cette grande distinction dans la finalité d'une évaluation, quelques auteurs ont désiré en ajouter deux autres. D'un côté, l'évaluation serait aussi utile pour améliorer l'état des connaissances à propos de l'efficacité d'une intervention, pour construire une théorie et pour publier des articles scientifiques (Patton, 1997). Elle se rapproche ainsi de la pratique de recherche. D'un autre côté, la finalité de l'évaluation de programmes ne serait pas uniquement liée aux programmes et à leurs composantes, mais aussi à l'amélioration des conditions sociales et collectives, ce que disait déjà Carol Weiss au début des années 1970. L'évaluation devient alors un moyen, elle n'est pas une fin en soi. Certains cherchent même à renforcer le pouvoir d'agir (*empowerment*) des participants au programme et à l'évaluation, ce dont nous traiterons au chapitre 15.

Planifier une évaluation

Si l'on fait abstraction de la concomitance des sous-processus, un processus évaluatif se déroule habituellement en deux phases, parfois successives mais qui peuvent être aussi parallèles ou en boucle. Il s'agit de la phase préévaluative et de la phase évaluative. Certains auteurs leur ajoutent une troisième phase, celle de l'utilisation des résultats des évaluations pour la prise de décisions (Grembowski, 2001). Notons cependant que tous les évaluateurs ne semblent pas être d'accord pour endosser cette responsabilité.

Le nom de la première phase est une traduction du concept anglophone d'*evaluability assessment*. L'étape préévaluative (Beaudry et Gauthier, 1992) permet de s'assurer, non seulement qu'il est possible de réaliser l'évaluation d'un programme, mais aussi que les évaluateurs disposent

de toutes les informations sur la demande d'évaluation. Wholey propose quant à lui deux idées à cet égard. Il distingue l'étude de la faisabilité de l'évaluation (la présence des conditions permettant de réaliser l'évaluation), d'une part, et l'analyse du caractère évaluable d'un programme (essentiellement liée à la cohérence du modèle logique, voir chapitre 5), d'autre part. Ainsi, les enjeux de l'évaluation, qui sont parfois politiques (chapitre 4), doivent être appréhendés, tout comme doivent être anticipés les conflits potentiels. Les tenants d'une évaluation centrée sur l'utilisation des résultats, comme Patton (1997) (chapitre 8), et ceux d'une évaluation réactive, comme Stake (2004), s'interrogeront immédiatement sur ce qui préoccupe les parties prenantes et sur la manière dont elles entendent mettre les recommandations en œuvre.

Ensuite, le cœur de cette démarche préévaluative est de vérifier que les individus concernés par l'évaluation savent précisément en quoi consiste le programme qu'il s'agit d'évaluer. Si nombre d'organisations mettent des interventions en œuvre et souhaitent qu'une appréciation soit faite sur ces dernières, il est rare qu'elles disposent d'une description détaillée du programme, de ses composantes et de ses objectifs. Notre expérience montre que cette lacune constitue la norme plus que l'exception. Aussi le travail fondamental à réaliser d'emblée est-il de reconstituer la logique du programme (chapitre 5) tel qu'il a effectivement été mis en œuvre. Même si cette logique est parfois présente sous diverses formes dans les documents de planification, il existe toujours une différence entre ce que l'on a prévu de faire et ce que l'on fait vraiment et le programme est souvent appliqué de façon différente d'un territoire à l'autre et d'une personne à l'autre. C'est ce que mesure l'évaluation d'implantation (Carroll, 2007). Il demeure donc indispensable de décrire de manière consensuelle le programme dont il est question dans l'évaluation ainsi que ses composantes. La place des parties prenantes dans ce processus est notamment abordée au chapitre 13. Si l'on ne sait pas exactement quels sont les effets attendus d'un programme, comment en évaluer l'efficacité? Si l'on ne connaît pas précisément l'enchaînement des activités et les éventuels liens de causalité avec des effets attendus, il est impossible d'attribuer au programme la responsabilité des changements observés.

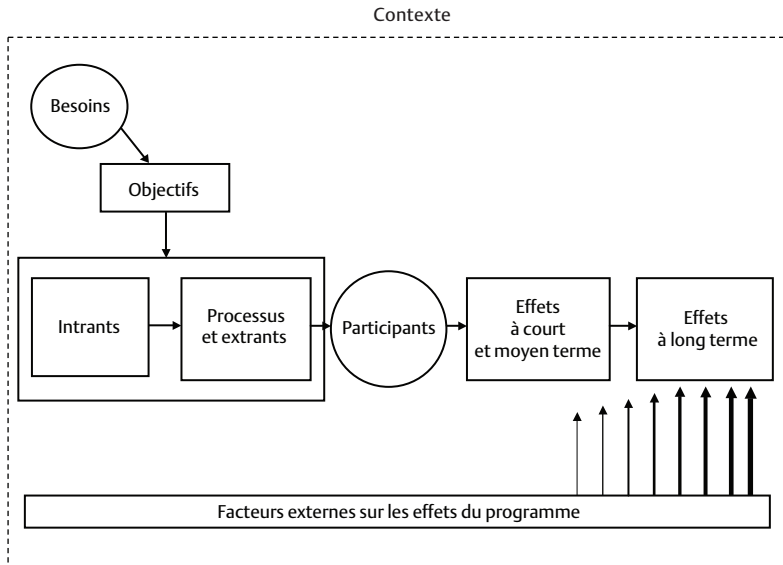
Selon les écoles de pensée et leur terminologie, il faudra donc disposer, à la fin de cette préparation, d'un modèle logique, d'une théorie ou de la logique du programme à évaluer. Certains auteurs établissent une diffé-

rence entre le modèle logique opérationnel d'un programme et son modèle théorique conceptuel. Dans la même veine, Chen (2005) différencie le modèle de changements (hypothèses descriptives à l'égard des relations causales qui expliquent l'atteinte des objectifs) du modèle d'action (hypothèses prescriptives à l'égard des composantes du programme qui vont faire en sorte qu'il fonctionne). Mais retenons surtout ici que la description précise d'un programme et des liens entre ses composantes est indispensable à tout processus évaluatif.

Pour y parvenir, il est souvent utile de faire une représentation graphique des relations entre les activités prévues et les effets attendus, dont on choisira la forme et les symboles. Les chapitres de la seconde partie de cet ouvrage fournissent des exemples intéressants à ce sujet. L'important, dans ces images, est que les relations soient claires et suffisamment précises pour aboutir à une compréhension commune du programme étudié et pour orienter la manière dont les questions d'évaluation seront posées.

Ces représentations peuvent également être utiles pour montrer que la logique d'un programme ne produira pas d'effets tant que, par exemple, la manière dont celui-ci est conçu ira à l'encontre des connaissances dont nous disposons sur les manières d'intervenir dans un domaine particulier. Par exemple, les données probantes actuelles en matière de réduction de la mortalité maternelle dans les pays du Sud nous apprennent que la formation d'accoucheuses traditionnelles et la tenue de consultations prénatales sont relativement inefficaces. Si la construction de la logique d'un programme par ses acteurs met en évidence qu'ils développent leur programme en suivant ces liens de causalité, l'évaluateur pourra d'emblée porter un jugement sur cette « logique illogique ». Dans un tel cas de figure, la phase préévaluative permet d'éviter la mise en œuvre d'une évaluation des effets inappropriée et de suggérer, par exemple, de réaliser une évaluation des possibilités d'action afin de revoir la logique du programme. Nous proposons en figure 1.3 une représentation graphique des éléments composant un programme, mais il en existe beaucoup d'autres.

Si l'on conserve l'exemple du programme de maternité à moindre risque, les besoins (mortalité, morbidité) des participantes (les femmes en âge de procréer) impliquent la mise en œuvre d'un programme qui vise des objectifs (réduire la mortalité maternelle) dans un contexte particulier (celui d'une région de l'Afrique). Ce programme est composé d'*intrants* (les ressources financières et humaines pour la mise en place des activités) et de *processus* (formation, chantier de construction) qui

FIGURE 1.3**Représentation générique de la logique d'une intervention**

produisent des *extrants* (livres, bâtiments de la maternité) dans le but de générer des effets à court terme (qualité et accessibilité aux soins) sur les participantes (parturientes et sages-femmes) ou des effets à plus long terme (réduction du taux de mortalité maternelle).

Dans la figure 1.3, les participants sont visuellement bien séparés des processus afin de souligner le fait qu'un programme se justifie sur la base des changements qu'il cherche à induire en faveur de ses bénéficiaires. Les flèches horizontales mettent l'accent sur la chaîne des résultats, les liens de causes à effets. L'utilisation de flèches de plus en plus grandes au-dessous des effets permet d'insister sur l'importance de rester modeste quant à l'atteinte des effets. Car plus on se dirige vers la droite de la figure (vers les effets à long terme), plus des facteurs externes (autres programmes, acteurs, changements contextuels) sont susceptibles d'être à l'origine de ces effets, et donc, moins on peut en attribuer le bénéfice au programme. Ce n'est évidemment pas un programme de santé publique, quelle que soit son ampleur, qui pourra contribuer seul à réduire la mortalité maternelle. Par exemple, on sait parfaitement que l'accroissement du taux de scolarisation des petites filles aura sur la mortalité maternelle un effet à long terme aussi important, sinon plus, qu'un programme de santé publique.

Une fois la logique de l'intervention circonscrite, il faut pouvoir déterminer les questions d'évaluation auxquelles il s'agira de répondre. Dans de nombreux contextes, les commanditaires d'une évaluation veulent tout savoir sur tout, souvent en peu de temps et avec un budget restreint. Il est donc souvent indispensable de sérier les demandes et de vérifier la faisabilité de l'évaluation. On pourra mettre à profit l'approche suggérée au chapitre 9 pour mieux cerner les questions d'évaluation à l'aide des critères de temps, de ressources et d'accès aux données, ainsi que des critères politiques ou stratégiques.

Nous avons mis cette grille en pratique en 2005, lors d'une évaluation réalisée à Haïti. Une fois bien compris les enjeux et les besoins, une liste de la totalité des questions d'évaluation qui permettraient de répondre à l'ensemble des requêtes formulées dans le mandat que nous avons reçu a été dressée. Devant l'ampleur de la tâche et l'impossibilité manifeste de répondre à toutes ces interrogations, une liste des questions prioritaires a été constituée. Deux séries de critères ont été successivement appliquées pour effectuer cette sélection. La première série était d'ordre technique, reposant sur trois indicateurs : le temps, le budget et l'accès aux données. Certaines questions ont aussi été immédiatement écartées pour des raisons méthodologiques. La seconde série de critères était d'ordre stratégique et liée à l'utilisation potentielle des résultats de cette évaluation.

Une fois réglées ces questions cruciales, il faut élaborer un plan d'évaluation. Des exemples sont fournis dans la seconde partie de cet ouvrage. Encore une fois, afin de ne pas créer de distance inutile entre les évaluateurs et les parties prenantes, il est parfois préférable d'employer le terme « plan » plutôt que « protocole » ou « devis » d'évaluation, qui connotent le monde de la recherche. Pour chacune des questions d'évaluation retenues, il faut déterminer :

- les données nécessaires pour y répondre ;
- les instruments de collecte de ces données ;
- les personnes qui détiennent ces données et les lieux où elles se trouvent ;
- le moment où les instruments de collecte seront utilisés et par qui ;
- les méthodes d'analyse des données.

Ensuite, une évaluation de programmes se gère comme toute action : il faut prévoir une liste des tâches, un calendrier d'exécution, une description des ressources humaines et un budget. Nous avons dit qu'une

évaluation était surtout utile pour la prise de décisions, et notamment pour améliorer les programmes ; par conséquent, il est indispensable, dès la planification, de s'interroger sur la manière dont seront reçus les résultats. Il faut réfléchir aux stratégies que l'on entendra mettre en œuvre pour communiquer ces résultats (présentations face à face, résumé vulgarisé, formats différents selon les parties prenantes, etc.) et faire en sorte que les personnes concernées puissent se les approprier. Une évaluation n'est réellement utile que si ses résultats sont utilisables et utilisés.

Les rares travaux empiriques traitant des conditions d'utilisation de l'évaluation (Cousins et Shulha, 2006) montrent que celles-ci s'apparentent aux conditions d'utilisation de la recherche. Ce riche domaine d'études montre qu'il existe une multitude de conditions qui favorisent l'utilisation des connaissances produites par les activités des chercheurs. Elles peuvent être classées en cinq catégories (Dagenais, 2010 ; Dagenais *et al.*, sous presse) : 1) l'opinion des utilisateurs potentiels, 2) leur expertise individuelle, 3) le contexte organisationnel propre aux utilisateurs, 4) les stratégies de soutien à l'utilisation des résultats et 5) les caractéristiques des chercheurs et de leur milieu. Le tableau 1.1 présente des exemples pour chacune de ces catégories.

TABLEAU 1.1**Catégories de conditions qui favorisent l'utilisation de la recherche**

Caractéristiques des utilisateurs	Réceptivité et attitude positive face à la recherche Utilité perçue de la recherche Expertise par rapport aux connaissances issues de la recherche
Contexte organisationnel	Culture organisationnelle valorisant la recherche Niveau d'engagement de l'organisation dans le processus de transfert Leadership fort de la direction Consensus concernant la nature des besoins en matière de connaissances Vision commune et partagée des résultats à atteindre Ressources dédiées pour les activités de TC
Caractéristiques des connaissances	Adéquation avec les valeurs et les besoins des utilisateurs Applicabilité Niveau de participation des utilisateurs à la production des connaissances Accessibilité des informations Délais de production appropriés
Stratégies de transfert et de soutien	Doivent tenir compte des caractéristiques des groupes cibles et de leurs besoins Doivent être basées sur une relation de confiance avec les utilisateurs Mécanismes d'échange (formels et informels) Langage commun Activités menées au moment opportun Adaptation du format des connaissances Accompagnement et suivi régulier (mesure systématique des progrès)

La mise en œuvre d'une évaluation : types et approches

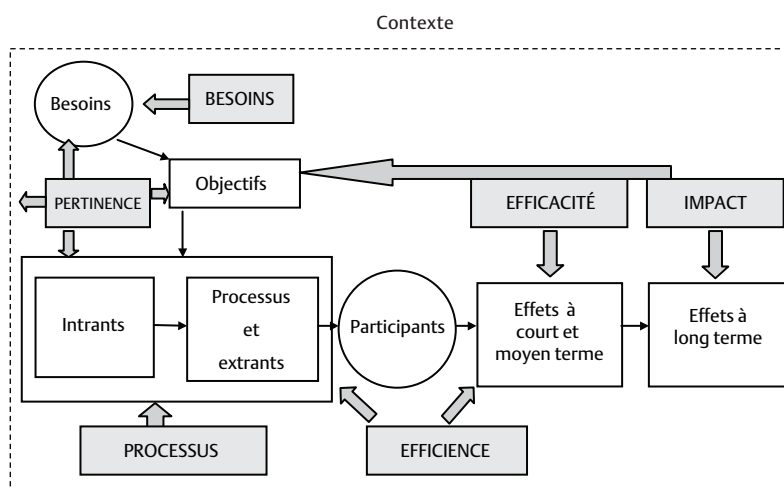
Nous avons choisi de négliger ici le côté théorique de la mise en œuvre d'une évaluation. En effet, plutôt que de discourir sur la réalisation, nous avons préféré présenter au lecteur des études de cas concrètes offrant des réponses à une question que les étudiants posent souvent : « comment fait-on l'évaluation d'un programme ? » Le lecteur trouvera donc dans la seconde partie de cet ouvrage plusieurs illustrations de démarches évaluatives. Pour accroître l'étendue de ces exemples, nous présenterons des évaluations effectuées dans plusieurs domaines comme l'action sociale, l'éducation, la santé ou l'administration, évaluations de différents types et qui s'appuient sur des approches distinctes organisées dans des contextes sociaux variés.

Les types d'évaluation

La représentation schématique de la figure 1.4 est une introduction efficace aux différents types d'évaluation définis comme l'objet sur lequel porte l'appréciation de l'évaluateur. Dans le cas de l'évaluation d'un programme, cet objet peut être une composante du programme ou encore la

FIGURE 1.4

Les types d'évaluation et leurs relations avec les composantes d'un programme



relation entre ses composantes. Il serait trop long ici de passer en revue les différents types d'évaluation, aussi en donnerons-nous simplement quelques exemples.

Le tableau 1.2 propose quelques exemples de questions choisis en fonction des divers types d'évaluation. Les études de cas de la seconde partie de cet ouvrage fournissent plusieurs illustrations empiriques de ces questions génériques.

TABEAU 1.2

Types d'évaluation et questions génériques

Types d'évaluation	Exemples de questions génériques
Évaluation des besoins	Quels besoins le programme vise-t-il à satisfaire? Les besoins des participants ont-ils évolué depuis le début de la mise en œuvre? Voir chapitre 16.
Évaluation de la pertinence	Le programme répond-il aux besoins des participants? Le programme prend-il en compte le contexte de sa mise en œuvre? Voir chapitre 14.
Évaluation du processus	Le programme est-il mis en œuvre tel qu'il a été planifié? Quels facteurs facilitent ou entravent la mise en œuvre du programme? Voir chapitres 17 et 19.
Évaluation de l'efficacité	Les effets constatés à court terme correspondent-ils aux objectifs initialement fixés? Comment les effets relevés chez les participants au programme se comparent-ils avec la situation des non-participants? Voir chapitres 15, 16, 18, 19 et 20.
Évaluation de l'impact	Les effets constatés à long terme correspondent-ils aux objectifs initialement fixés?
Évaluation de l'efficience	Quel est le rapport entre les ressources et les effets du programme? Ce programme est-il plus onéreux qu'un autre qui produirait le même effet? Voir chapitre 11.

Notons que de nombreuses autres typologies sont proposées dans les écrits sur l'évaluation. Mais peu importe celle qui est privilégiée, il est toujours essentiel de bien analyser le contexte : selon sa nature, un programme n'aura pas les mêmes effets et ne sera pas mis en œuvre de la même manière. Il faut aussi rappeler que des effets non prévus, qu'ils soient favorables ou non aux participants, peuvent être générés (Morell, 2010), et ils doivent impérativement être relevés.

Chen (2005) estime que certains types d'évaluation s'appliquent mieux à certains moments de la vie d'un programme. Mais d'autres types d'évaluation peuvent aussi être employés. Par exemple, lorsque l'on souhaite concentrer les ressources sur des activités qui ont plus de chances de

réussir, il est parfois utile de réaliser une évaluation des *possibilités d'action*. L'évaluation de la *cohérence* des activités d'un programme est aussi quelquefois requise. Certains distinguent l'évaluation de l'*implantation* (décrire ce qui se passe) de l'évaluation des processus (comprendre la dynamique interne). Au sein de la catégorie de l'évaluation d'efficacité, aussi nommée « évaluation économique » (chapitre 11), il est utile de distinguer quatre types particuliers d'évaluation : minimisation des coûts, coût-efficacité, coût-utilité et coût-bénéfices (Drummond *et al.*, 1998). Il ne faut pas oublier, enfin, la possibilité de réaliser une évaluation de l'équité (Potvin *et al.*, 2008 ; Ridde *et al.*, 2009 ; Ridde *et al.*, 2010), ce qui est encore trop rarement fait.

Les approches évaluatives

Pour répondre aux questions, l'évaluateur peut avoir recours à plusieurs approches évaluatives. Le terme d'approche renvoie ici à la manière dont les experts interagissent avec les individus concernés par l'évaluation. On peut aussi définir l'approche comme étant la façon dont l'évaluateur assume son rôle. L'approche utilisée varie en fonction des situations, de l'organisation dans laquelle l'évaluateur évolue, de ses caractéristiques personnelles et des besoins précisés par les personnes concernées. Il n'y a pas d'approche qui soit *a priori* meilleure qu'une autre. Il s'agit de choisir celle qui paraît la plus pertinente dans le cas étudié. En forçant un peu le trait de l'idéal-type wébérien, il est possible de mettre en évidence les éléments fondamentaux qui caractérisent trois approches particulières.

L'évaluateur qui emploie une *approche directive* a un rôle d'expert et reste relativement neutre et distant par rapport à l'objet évalué. Il dirige l'ensemble des opérations et prend les décisions. Les différents acteurs ne sont impliqués qu'à titre de source d'information, et cela, à des moments particuliers de la démarche. Seul l'évaluateur a l'autorité pour juger de la valeur et du mérite du programme.

L'*approche axée sur le développement* d'un programme est moins connue et d'un emploi moins répandu (Patton, 1994 ; Ridde *et al.*, 2006). Son intérêt est que les résultats et le processus même de l'évaluation sont utilisés pour faire évoluer le programme au cours de sa mise en œuvre. L'expert en évaluation guide les parties prenantes pour faire en sorte que le projet s'améliore, s'adapte et réponde mieux aux besoins des partici-

pants. Il est membre à part entière de l'équipe du projet et participe aux prises de décisions, tant sur le plan de l'évaluation que sur celui du développement du projet.

Quant à l'*approche participative*, elle préconise la mise en œuvre du processus évaluatif selon une démarche de coopération entre des évaluateurs et des personnes concernées par l'évaluation, mais qui ne sont pas des évaluateurs professionnels. Trois raisons peuvent justifier le recours à une telle approche : la première est d'ordre pragmatique (accroître les chances de réussite du processus et maximiser l'utilisation des résultats), la deuxième est politique (souci démocratique de participation citoyenne et d'émancipation), et la troisième est épistémologique (il n'existe pas de réalité unique et la pluralité des points de vue est essentielle à la construction d'un argument) (Weaver et Cousins, 2005 ; chapitre 11). On peut distinguer deux tendances au sein de l'approche participative :

- L'*évaluation pratique* sert à résoudre des problèmes, à améliorer un programme et à prendre des décisions. Elle est réalisée dans le cadre d'un partenariat égalitaire entre l'évaluateur et les parties prenantes, ces dernières étant essentiellement les personnes devant prendre les décisions.
- L'*évaluation émancipatrice* est un moyen, mais aussi, dans une certaine mesure, une fin en soi. Le processus évaluatif doit favoriser l'accroissement du pouvoir d'agir (*empowerment*) des parties prenantes, dans une perspective de justice sociale. Les décisions à l'égard du processus évaluatif sont exclusivement prises par les parties prenantes, et non par l'évaluateur (Cousins et Whitmore, 1998 ; Ridde, 2006).

Les normes de pratique

À la suite de la création de multiples associations professionnelles et regroupements nationaux se consacrant à l'évaluation de programmes, essentiellement dans les pays à revenus élevés, l'année 2003 a vu naître l'Organisation internationale de coopération en évaluation (OICÉ). Dans les pays à moyens et faibles revenus, de telles associations sont encore rares. Dans l'Afrique de l'Ouest francophone, par exemple, l'une des premières à voir le jour a été le Réseau nigérien de suivi et évaluation (ReNSE), créé dans la foulée de la première conférence de l'Association africaine d'évaluation (AfrEA), tenue en 1999. L'une des premières tâches

de ces associations professionnelles d'évaluateurs a été de se doter de normes de pratique, de codes d'éthique ou, plus récemment, d'une liste des compétences essentielles propres à cette profession (voir chapitre 12). L'Association américaine d'évaluation, par exemple, a produit une liste de 30 critères garantissant la qualité d'une évaluation, critères qui sont répertoriés selon quatre dimensions clés d'une évaluation de programmes de qualité : l'utilité, la faisabilité, la propriété et la précision. L'AfrEA a tenté d'adapter ces principes (Rouge, 2004). Les Suisses ont remplacé la troisième dimension par la déontologie. Les Français se sont munis de sept principes directeurs : pluralité, distanciation, compétence, respect des personnes, transparence, opportunité et responsabilité. Enfin, depuis peu, tant aux États-Unis qu'au Québec, des tentatives ont cours pour établir une liste des compétences essentielles que devraient détenir les évaluateurs (Doré et Marceau, 2006 ; Stevahn *et al.*, 2005 ; voir aussi en encadré, p. 221). Nous avons testé l'utilisation de compétences en Afrique (Ridde *et al.*, 2009). La Société canadienne d'évaluation dispose depuis 2010 d'un référentiel de compétences et d'un processus d'accréditation volontaire des évaluateurs. Une centaine de personnes étaient accréditées au début de 2012.

L'objectif de cette introduction n'était pas de dresser un portrait exhaustif du champ de l'évaluation de programmes. En effet, il existe de multiples manières d'approcher ce champ de pratique et de définir les termes qui y sont employés. Nous espérons cependant avoir initié le lecteur aux concepts essentiels à la compréhension de l'évaluation de programmes. Les chapitres de la première partie de l'ouvrage approfondissent certaines notions abordées nécessairement trop rapidement ici. Dans la seconde partie, des études de cas détaillées illustrent des applications empiriques de ces concepts et montrent notamment la richesse et la diversité des pratiques évaluatives.

2

Recherche et évaluation de programmes

Miri Levin-Rozalis

Il semble que les liens et les différences entre l'évaluation et la recherche suscitent constamment des questions. À la suite des congrès annuels de l'European Evaluation Society et de l'American Evaluation Association de 2002, j'ai décidé de relever le défi que constitue toute tentative de clarification des différences existant entre ces deux secteurs d'activités. Ma vision de l'évaluation, en tant que service aux intervenants et outil pour améliorer la conduite d'un projet, a fortement influencé la teneur de ce chapitre.

Dans mon travail d'évaluatrice, j'ai souvent été témoin du manque de connaissances concernant ce qu'est l'essence même de l'évaluation, et en particulier de la différence entre l'évaluation et la recherche. Dans la littérature, on rencontre souvent l'expression « recherche-évaluative » comme s'il s'agissait d'une sorte d'hybride, mais dans bien des cas, il s'agit du résultat d'un métissage qui ne correspond ni à une recherche adéquate ni à une évaluation appropriée.

Dans ce chapitre, je traiterai des similarités et des différences qui distinguent l'évaluation et la recherche. Je veux démontrer que ce sont des

Ce chapitre est adapté d'un article paru en anglais, en 2003, dans la *Revue canadienne d'évaluation de programmes* (vol. 18, n° 3).

disciplines bien différentes, et cela, malgré les similarités qu'elles présentent et qui proviennent du fait qu'elles partagent des concepts, des instruments et, dans certains cas, des méthodes. Selon moi, la difficulté de faire la distinction entre la recherche et l'évaluation cause un préjudice à cette dernière. Les efforts constamment consentis par les évaluateurs pour appliquer à l'évaluation les critères de la recherche les forcent à abandonner les particularités de l'évaluation. Ce faisant, le caractère unique et la qualité de l'évaluation se trouvent diminués.

Les débats philosophiques qui ont cours sur la recherche scientifique portent généralement sur son essence – rationalistes contre empiristes, positivistes contre tenants de l'interactionnisme symbolique – et tendent à mettre l'accent sur des aspects particuliers de cette entreprise (Bechler, 1987; Caws, 1965; Copi, 1961; Hempel, 1965; Kincheloe et McLaren, 2000; Kuhn, 1962; Popper, 1959). Cependant, plusieurs critères et définitions sont reconnus par tous. On convient généralement que le but de la recherche, en tant que méthode scientifique, est de découvrir et de comprendre les lois fondamentales selon lesquelles le monde fonctionne, mais aussi d'accroître l'ensemble des connaissances afin d'arriver à une meilleure maîtrise de la nature et de satisfaire le désir naturel des êtres humains d'acquérir de nouvelles connaissances. Ainsi, dans le secteur de la recherche, recueillir des connaissances est le premier but de toute activité; la connaissance pure ayant une valeur intrinsèque.

On n'a pas proprement déterminé l'essence de l'évaluation comme on l'a fait pour la recherche scientifique. Cependant, par définition, l'évaluation est un processus constant de changement et de développement, car il est utilisé dans un objectif d'investigation ou d'étude. On considère généralement l'évaluation comme un outil pour les gestionnaires de programmes, les décideurs et les intervenants. On l'emploie dans plusieurs domaines, allant des processus éducatifs et sociaux aux services de santé et à l'agriculture, et elle peut se faire, par exemple, à l'aide d'études strictement quantitatives ou totalement non directives. L'évaluation a aussi de très nombreuses fonctions. D'abord, elle peut chercher à déterminer le degré de réussite d'un programme, à voir si un programme donné a évolué ou atteint ses buts, à connaître les effets secondaires des activités afin d'améliorer l'efficacité d'un projet éducatif, communautaire ou social en cours ou de contribuer à sa poursuite (Cronbach, 1963, 1983; House, 1980; JCSEE, 1981; Nevo, 1989; Patton, 1981; Scriven, 1967; Stufflebeam, 1971;

Tyler, 1950). L'évaluation peut aussi être une « opération pratique, matérielle et politique dont l'intérêt est d'examiner et d'améliorer nos façons de porter des jugements interprétatifs sur la valeur des actions humaines qui se déroulent dans des contextes sociaux et culturels particuliers » (Schwandt, 2002, p. 2).

Les demandes faites aux évaluateurs et à l'évaluation en général ont évolué, passant de l'examen d'objectifs opérationnels mesurables, dans les années 1950, à la cueillette, dans les années 1970, d'informations utiles pour les décideurs, allant même jusqu'à façonner l'intervention elle-même (Nevo, 1989; Scriven, 1967; Stufflebeam *et al.*, 1974). À la même époque, il est devenu évident que les évaluateurs, pour répondre efficacement aux besoins des différents projets, devaient être inventifs (Patton, 1981). Dans les années 1980, on attendait des évaluateurs qu'ils puissent tenir compte de tous ceux qui pouvaient être touchés par l'activité d'évaluation et utiliser l'évaluation pour restructurer et recadrer les concepts et points de vue pour leurs clients, les intervenants, décideurs et usagers du projet (Abma, 1997; Guba et Lincoln, 1989; Pawson 1996), ou encore influencer la structure du pouvoir en outillant ceux qui faisaient l'objet de l'évaluation (Brisolara, 1998; Cousins et Whitmore, 1998; Fetterman, 1996).

Ces changements et ces demandes entraient en conflit avec les cadres définis et structurés de la recherche. On attendait de l'évaluateur qu'il soit à même de déterminer quelles connaissances étaient importantes et pour qui, de savoir comment recueillir, traiter et interpréter ces connaissances, de décider des conclusions à tirer et de donner des conseils qui se fondent sur ces conclusions. Étant donné ces exigences, les évaluateurs se retrouvaient souvent pris entre le marteau et l'enclume. D'un côté, on leur demandait de trouver les réponses à un grand nombre de questions diversifiées, et de l'autre, il leur manquait un cadre clair et des outils appropriés pour le faire, tout en procédant dans un environnement en évolution constante (Eyken *et al.*, 1995; Guba et Lincoln, 1989; Tucker et Dempsey, 1991). Certains évaluateurs, pour régler ces problèmes, ont emprunté des méthodes de diverses disciplines (Finne *et al.*, 1995; Kazi, 1996; Tucker et Dempsey, 1991) alors que d'autres ont établi des liens avec la théorie (Chen, 1990; Chen et Rossi, 1981, 1992; Pawson et Tilley, 1997).

Malgré les différences entre l'évaluation et la recherche et celles qui existent entre les demandes adressées aux évaluateurs et les mandats attribués aux chercheurs, une confusion concernant le rôle de chacun

persiste, et cela pour plusieurs raisons. Parmi elles, la définition même de la recherche, qui n'est pas claire et, particulièrement dans le domaine des sciences sociales, qui évolue constamment. La même situation prévaut en évaluation. Devant des conditions souvent embrouillées, les évaluateurs eux-mêmes cherchent un cadre de référence bien défini et reconnu avec lequel travailler. Les procédures et critères de la recherche fournissent ces cadres de références reconnus, et la recherche continue d'être tenue en plus haute estime que l'évaluation. Ces assertions, résumées dans le tableau suivant, seront approfondies dans les prochaines pages.

TABLEAU 2.1
Points communs et différences entre l'évaluation et la recherche

	Évaluation	Recherche
Domaines d'application	Application de l'examen aussi large que possible. Application limitée des résultats, qui sont concentrés sur le projet. Le but est de fournir de la rétroaction concrète.	Application de l'examen aussi large que possible. Application des résultats aussi large que possible. Le but est d'accroître l'ensemble des connaissances scientifiques.
Théorie	Dépendance envers le terrain : la théorie sert à augmenter la compréhension des résultats.	Dépendance envers la théorie : la théorie est à la fois la source et l'objectif.
Méthodologie	Le choix du cadre de l'évaluation et des méthodes de cueillette de données provient du terrain. L'évaluateur réagit.	Le choix du cadre de recherche et des méthodes de cueillette de données provient de la théorie. Le chercheur agit.
Généralisation	Effort pour comprendre ce qui se produit dans un projet en particulier.	Effort pour formuler une loi générale ; la validité externe est importante.
Pertinence	Utilité pour le projet.	Accroissement des connaissances scientifiques.
Causalité	Accent mis sur la validité interne ; ce qui en recherche constitue un artéfact est vu en évaluation comme une variable interne permettant d'obtenir une causalité.	La validité interne est importante ; accent sur un petit nombre de variables causales isolées des autres variables.

Les domaines d'application

La principale différence qui existe entre l'évaluation et la recherche réside dans leurs champs d'application. Les deux cherchent à augmenter notre compréhension dans divers domaines, mais le type de compréhension à

laquelle elles contribuent diffère. Le paradigme dominant en recherche vise l'expression de connaissances générales sous forme de lois ; ainsi, la compréhension qu'on veut atteindre est celle d'une loi abstraite et universelle pouvant s'appliquer au plus grand nombre d'événements possibles. (En sciences sociales, aujourd'hui, ce paradigme est remis en question par certains courants tels que l'approche interprétative, le relativisme, le poststructuralisme et la théorie critique, mais ces approches veulent également parvenir à une compréhension générale des phénomènes observés.) À l'opposé, le but de l'évaluation est de recueillir des connaissances pour développer une compréhension relativement à une activité concrète, un projet, et d'appliquer cette compréhension au projet lui-même sous forme de rétroaction. L'évaluation tente d'examiner les influences mutuelles d'un maximum de variables à un moment et à un endroit donnés.

Le but de la recherche est d'enrichir un ensemble de connaissances, la connaissance y ayant une valeur en soi, alors qu'en évaluation, la connaissance est un moyen d'arriver à certaines fins. Sa valeur réside d'abord et avant tout dans la rétroaction donnée au projet. Ainsi, l'évaluation est un processus circulaire dont le cycle débute et se termine avec le projet. Dans le cas d'interventions sociales, en éducation ou dans la communauté, les suivis effectués par les évaluateurs deviennent une nécessité. La capacité de se tenir à l'écart et d'observer, de percevoir et de comprendre les processus, de les organiser, de les analyser et de les rendre sous la forme d'une rétroaction intelligible est une contribution importante qui augmente l'efficacité d'un projet ou d'une activité dans des conditions qui évoluent.

À propos de l'application de l'évaluation et de la recherche, il est important de distinguer deux points : les procédures d'examen (questionnaires, entrevues, observations) que les évaluateurs et les chercheurs emploient pour recueillir des données, et les résultats et conclusions provenant des données, ainsi que leurs conséquences. La raison de cette distinction se trouve dans les principales différences entre l'évaluation et la recherche, dans leur application potentielle.

En ce qui concerne les procédures, une application étendue est importante, tant pour la recherche que pour l'évaluation. On recommande donc aux évaluateurs d'utiliser des procédures d'examen systématiques et répétables qui puissent être critiquées et faire l'objet d'un suivi. Si on utilise des procédures statistiques, il est important que ce soient des

méthodes reconnues. Si cela est possible, mais souvent ce ne l'est pas, il est préférable d'employer des instruments connus pouvant être utilisés dans d'autres cadres.

On constate une autre différence entre l'évaluation et la recherche lorsqu'on aborde la question des résultats. Dans tout type de recherche, on souhaite que l'application des résultats et des conclusions qui en sont tirées soit la plus large possible afin de servir de base à des lois universelles. C'est le premier devoir de la recherche.

En évaluation, les résultats ne visent à être pertinents que pour le projet évalué, tout comme ils ne seront exacts que pour une certaine période du projet. On ne cherche pas à obtenir des lois immuables pouvant être appliquées au-delà du projet à l'étude. Cela ne signifie pas, comme le dit Pawson (1996), que l'on ne puisse considérer les « légendes » et les « histoires » provenant du milieu, sans conceptualisation ni apprentissage (Patton, 1997), comme de véritables résultats. Ceux-ci ne font que confirmer aux membres du projet la compréhension de leur situation qu'ils possédaient déjà avant l'évaluation. On ne prétend pas par là qu'il est impossible d'obtenir des résultats ou, plus souvent, des conclusions qui aient une valeur dépassant le cadre du projet lui-même, cela arrive fréquemment, mais ce n'est pas le but premier de l'évaluation. J'aborderai la question de la généralisation plus loin dans ce chapitre.

Lorsque des évaluateurs utilisent des procédures de recherche pour réaliser une évaluation, ils emploient des instruments adaptés aux questions auxquelles ils s'intéressent ; il en résulte une plus large applicabilité des résultats. Paradoxalement, des variables définies, qui restent générales et abstraites, peuvent accroître l'applicabilité des résultats, mais elles diminuent l'utilité des conclusions pour un projet. C'est-à-dire que la possibilité de généraliser les résultats et les variables augmente, mais celle de donner une rétroaction intelligible et cohérente sur la situation du projet dans son ensemble en est de beaucoup réduite. Ainsi, les instruments propres au domaine de la recherche donnent des réponses abstraites qui ont un degré élevé de généralisation et qui sont appropriées pour la recherche, des réponses valides qui sont recueillies à l'aide d'instruments fiables et qui peuvent être répliquées. La qualité de l'évaluation en est cependant amoindrie, car les gestionnaires du projet ne reçoivent pas de réponses utiles à leur propre travail ; des réponses qui soient directement reliées aux différentes activités, problématiques et clientèles du projet. Les

résultats ayant un degré élevé de généralisabilité, constitués de variables théoriques et exprimés en langage scientifique sont difficilement traduisibles en propositions opérationnelles visant une amélioration. C'est pourquoi ils n'ont que peu ou pas de valeur en termes de rétroaction. (Un tel rapport peut être nécessaire pour justifier l'existence ou la poursuite d'un projet, ou encore pour obtenir du financement, mais non pour son fonctionnement quotidien.)

On retrouve en fait le même problème en recherche, où une application étendue des résultats s'oppose à une compréhension approfondie du sujet de recherche (Geertz, 1973; Goetz et LeCompte, 1984; LeCompte et Preissle, 1993; Sabar, 1990; Tucker et Dempsey, 1991). On élabore constamment de nouvelles méthodes de recherche comme l'étude de cas, l'ethnographie et autres méthodes de recherche qualitative (Creswell, 1998; Sabar, 1990) pour surmonter ces difficultés. Par ces approches, les chercheurs renoncent à une application trop étendue de leurs résultats afin de mieux comprendre les complexités de leur sujet de recherche. Dans de telles circonstances, les méthodes de recherche ressemblent aux approches utilisées dans l'évaluation de projets, sans toutefois en être identiques, puisque les buts et fonctions de la recherche qualitative ou ethnographique, ou encore de l'étude de cas, n'ont pas changé; la recherche visant toujours à enrichir l'ensemble des connaissances scientifiques en formulant des lois universelles, ainsi que l'application la plus étendue possible de ses résultats.

Les rôles de la théorie, des hypothèses et de la méthodologie

En recherche, la logique peut prendre trois voies pour relier la théorie et les données: la logique déductive, la logique inductive et l'inférence abductive. Dans la logique déductive, un lien solide existe entre les hypothèses de recherche et un postulat théorique antérieur. L'hypothèse – l'*explanandum* – est expliquée par les prémisses provenant de la théorie, et rien de nouveau ne s'y retrouve. Les postulats théoriques *a priori* constituent les *explanans*; ce sont eux qui expliquent l'hypothèse émise. La solidité du lien entre *explanans* (postulats *a priori*) et *explanandum* (hypothèse) n'est influencée par aucune information, établie ou nouvellement découverte et à confirmer. Cette façon de formuler des hypothèses est appropriée en recherche lorsqu'on examine une théorie ou qu'on essaie de la réfuter. Pour le chercheur, cela permet de s'assurer qu'il ne déviara pas en appliquant la

théorie en question à son sujet de recherche. Ainsi, les phénomènes observés sur le terrain ne sont pas assujettis à la logique déductive, le terrain n'étant que l'environnement dans lequel les hypothèses *a priori* peuvent être examinées (Copi, 1961; Copi et Burgess-Jackson, 1995).

La logique déductive est à l'opposé de la logique utilisée en évaluation, car celle-ci examine le contexte afin de découvrir les variables et les éléments en jeu, ainsi que les liens entre eux; le contexte ne sert pas à valider des variables et des hypothèses provenant d'une théorie préexistante. Pawson et Tilley sont les principaux tenants de la logique déductive. Voici un exemple emprunté de Weiss (1996) et cité par Owens et Rogers (1999). Lorsqu'un counseling en faveur de la contraception est associé à une réduction du nombre de grossesses, il apparaît que la cause de ce changement est le counseling. Mais qu'est-ce qui, dans cette démarche, a réellement causé le changement? Les connaissances fournies? Il se peut que le counseling aide à surmonter les tabous culturels qui vont à l'encontre de la planification familiale; il peut donner aux femmes une plus grande confiance en elles ainsi qu'une plus grande capacité de s'affirmer lors des relations sexuelles; il peut provoquer un déplacement du pouvoir dans les relations entre hommes et femmes. Ces raisons, ou d'autres encore d'ordres cognitif, affectif ou social, peuvent constituer les mécanismes qui entraîneront les résultats souhaités. Nous ne le saurons jamais en utilisant la logique déductive, car la théorie nous ramène, selon Pawson et Tilley (1997), aux mécanismes, aux contextes et aux résultats qui font partie de notre propre cadre de référence, ce qui n'est pas important lorsqu'on examine une théorie, mais qui est crucial lorsqu'on cherche à savoir ce qui fonctionne efficacement dans un projet.

En logique inductive, les hypothèses sont émises à partir des phénomènes généralisés observés sur le terrain. Ces hypothèses examinent les chances que des phénomènes se répètent pour formuler une loi générale des probabilités. Pour y arriver, on doit connaître les caractéristiques examinées du groupe auquel on s'intéresse et leurs conditions *a priori*. On peut faire une analogie avec la connaissance du fait qu'une pièce de monnaie a deux faces, et qu'elle retombe sur une seule de ses faces lorsqu'on la lance en l'air (Copi, 1961).

On ne retrouve généralement pas ces conditions en évaluation. Tout d'abord, il faut une généralisation empirique lorsque, de façon raisonnable, on croit se trouver devant un phénomène. Pour que cette généra-

lisation soit empirique, il faut préalablement connaître les caractéristiques de ce phénomène, et que ces caractéristiques aient déjà été examinées. Une partie de l'évaluation débute tôt, au cours de cet examen, ou même avant celui-ci. Elle doit d'abord tenter d'exposer le phénomène significatif, et ne peut donc examiner dès lors ses caractéristiques ou la probabilité de son apparition. Ensuite, le but d'un examen empirique est de généraliser au-delà d'un cas individuel, de tenter de formuler une loi universelle, ce qui n'est pas l'objectif principal de l'évaluation, pour toutes les raisons mentionnées précédemment.

L'évaluation ne dépend pas de la théorie, mais bien de l'action sur le terrain. Il existe bien sûr des théories de l'évaluation, des modèles d'évaluation et diverses approches pour les questions d'évaluation. On utilise diverses théories dans l'interprétation des résultats d'une évaluation, mais l'élément essentiel d'une évaluation est ce qui est recueilli sur le terrain, en ce sens que les questions posées par une évaluation ne proviennent pas d'une théorie. Elles proviennent plutôt du projet évalué, de ses objectifs, de son environnement et de ses usagers, ainsi que des besoins des personnes qui gèrent l'intervention, de celles qui prennent les décisions et des parties concernées à différents niveaux, et enfin des besoins des évaluateurs.

Ainsi, le travail d'un évaluateur ressemble à celui d'un médecin ou d'un psychologue qui examine un patient en considérant l'histoire et les conditions qui lui sont particulières. Médecins et psychologues trouvent par la suite une validation de leurs résultats et de leurs hypothèses dans la littérature existante et dans les théories pertinentes. De la même façon, les évaluateurs fondent leur travail sur un large éventail de connaissances pratiques et théoriques dans les divers domaines où ils interviennent. Lorsqu'un médecin traite un patient, il ne cherche pas à valider une théorie (à moins de faire de la recherche médicale) ; pourtant, il compte sur des connaissances obtenues grâce à la recherche. De la même manière, tout en se basant sur des connaissances théoriques, les évaluateurs ne cherchent pas à valider une théorie. Lorsqu'on tente de formuler des hypothèses à partir d'un cadre de référence théorique particulier, on limite la portée de l'évaluation et on empêche l'évaluateur d'émettre des hypothèses ne provenant pas de ce cadre. Le cadre de référence dicte les concepts à utiliser et la relation attendue entre ceux-ci. Dans une évaluation, une théorie convenant au projet peut être choisie à une étape ultérieure, au moment où l'évaluateur explique les résultats et tire des conclusions.

Le troisième type de recherche logique, l'inférence abductive, est un guide précieux dans l'évaluation (Peirce, 1960). Les principes de l'inférence abductive sont basés sur l'idée qu'il n'y a pas d'hypothèses *a priori*, pas de présupposés et pas de théories élaborées antérieurement. Chaque événement est étudié en lui-même et son importance propre est examinée (Shank et Cunningham, 1996). Des hypothèses concernant cet événement sont alors émises : cet événement est-il lié à d'autres et, si oui, comment ? Est-ce un événement isolé et, si oui, quelle est sa signification ?

Les explications que l'on attribue à ces événements sont des « hypothèses en probation ». Un processus cyclique de vérification des observations faites se met en place, ce qui permet d'approfondir et de modifier les explications (Levin-Rozalis, 2000).

Je ne prétends pas que l'évaluateur examine tous les faits et tous les événements qui se produisent sur le terrain. Un tel examen n'est ni possible ni souhaitable. Je soutiens seulement l'idée qu'en recherche scientifique, la théorie constitue le cadre de référence qui établit quelles données sont pertinentes et lesquelles ne le sont pas, alors qu'en évaluation, ce sont les besoins du projet et la réalité dans laquelle il est mis en œuvre qui déterminent les questions et les variables. On recommande que les gestionnaires du projet, de concert avec les évaluateurs, définissent la problématique. Mais même lorsqu'on ne procède pas ainsi, les évaluateurs prennent la situation sur le terrain en considération, et c'est cette situation, dans toute sa complexité, qui génère les questions qui seront explorées.

Après avoir remarqué un phénomène singulier dans un projet, il peut être intéressant et important de l'examiner selon des méthodes de recherche contrôlées. On peut ainsi produire de nouvelles connaissances propres à contribuer, non seulement au projet, mais également à l'ensemble des connaissances scientifiques.

Les évaluateurs exposent la réalité à l'aide de tous les moyens scientifiques dont ils disposent, et de la façon la plus systématique qui soit pour chaque projet en particulier. Les conditions que l'on retrouve dans le milieu faisant l'objet de l'investigation prescrivent les instruments qui seront utilisés tout autant que le langage employé. Par exemple, un questionnaire n'est pas indiqué pour une population ayant des difficultés de lecture. Ces conditions dictent également les concepts utilisés au cours de l'examen ; par exemple, utiliser les concepts et le langage du projet plutôt que ceux du jargon spécialisé de l'évaluateur.

Les évaluateurs organisent leurs résultats en modèles significatifs et, si possible, comparables. Il est possible de généraliser des résultats spécifiques dans des structures et des processus plus étendus, ou d'établir une théorie explicative, mais trouver une explication cohérente pour ce qui se produit dans un projet, incluant les succès et les échecs, demeure le but principal de l'évaluation.

Par exemple, le gestionnaire d'un projet évalué avait un style de gestion autoritaire qui interdisait toute initiative de la part de ses employés. Sachant cela, il devint possible de trouver une explication générale soutenue par les théories sur les liens entre autoritarisme et initiative. Mais nous ne pouvions postuler à l'avance que ce gestionnaire était autoritaire, puisque nous n'avions aucune idée de son style de gestion ni du type d'interactions qu'il avait avec les employés du projet, et nous ne savions pas si les questions concernant les relations direction-employés allaient être pertinentes. Les évaluateurs doivent avoir suffisamment de connaissances sur les styles de gestion, ainsi que sur l'autoritarisme et l'initiative, pour élaborer une explication pertinente. Si ce n'est pas le cas, ils peuvent chercher de l'aide auprès de professionnels qui connaissent les styles de gestion.

Les questions qui guident l'activité d'évaluation n'émanent pas d'une théorie, mais, dans de nombreux cas, le projet lui-même est une démonstration de l'application d'une théorie. Les programmes et méthodes du projet sont basés sur la logique d'une certaine théorie ou sur la manière dont les initiateurs, les gestionnaires et les bailleurs de fonds voient le monde. Parmi les éléments que l'évaluation examine, il y a la qualité de la logique déterminant les activités du projet, son lien avec ce qui se produit dans le milieu et la vérification de la congruence des résultats avec la logique (la théorie directrice). Dans ce cas, les questions que pose l'évaluateur font partie du dialogue qui s'engage avec la théorie.

La méthodologie traditionnelle de la recherche provient de la théorie ainsi que de ses hypothèses et généralisations empiriques. Elle est le fruit de la planification rigoureuse d'une stratégie d'opérationnalisation des hypothèses, c'est-à-dire de la formulation des variables en éléments observables. Un élément observable doit être ancré dans la réalité; une situation propice à l'observation doit être repérée ou créée artificiellement avant que les observations scientifiques nécessaires à la vérification de l'hypo-

thèse ne se déroulent. Pour cela, le chercheur doit choisir les variables les plus appropriées aux concepts de l'hypothèse: les sujets, le terrain et le cadre qui conviennent le mieux.

La méthodologie de l'évaluation diffère de celle de la recherche. Alors que les chercheurs peuvent choisir leurs propres sujets de recherche et proposer eux-mêmes une approche (nous l'appellerons la méthode active), les évaluateurs sont réactifs et proactifs. Ils réagissent aux événements et tentent parfois de les diriger. Les évaluateurs ne peuvent choisir eux-mêmes les participants au projet, ni son environnement ou ses variables. Le milieu et les participants sont prédéterminés et les variables, en partie à tout le moins, ne sont pas connues à l'avance. Il y a bien une idée générale des questions d'évaluation, mais celles-ci ne sont pas définies sous forme d'hypothèse et les variables ne sont pas opérationnelles. Parfois, même la définition nominale des concepts est matière à évaluation. Les instruments dont on se sert pour l'évaluation (entrevues, discussions, observations, questionnaires, vidéos, analyses de protocoles, analyses de dialogues, etc.) sont choisis et planifiés en fonction de la population concernée, des activités à vérifier, de la question à étudier, du temps et de l'argent dont on dispose, ainsi que du contrat entre les gestionnaires ou les initiateurs du projet et l'équipe d'évaluation.

La validité des résultats d'évaluation

Chaque projet étant différent des autres et exigeant également une approche distincte, il n'est pas rare d'avoir à traiter avec les questions de la validité des résultats et de la possibilité de les généraliser. En ce qui concerne la validité d'un cadre de recherche ou d'évaluation, on dit généralement que la recherche et le chercheur insistent sur la validité interne et externe (Cook et Campbell, 1976, 1979), alors que les évaluateurs mettent l'accent sur la validité interne seulement, puisqu'ils s'intéressent au fonctionnement immédiat d'un projet en particulier.

La méthodologie de l'évaluation surmonte avec assez de facilité la plupart des problèmes qui menacent la validité interne provenant des variables qui font l'objet de l'examen (par exemple, l'histoire, la maturation, la sélection ou la mortalité). Tous ces éléments sont inclus dans le champ de l'évaluation. Contrairement à la recherche, l'évaluation s'occupe d'un nombre considérable de variables, et non pas de variables isolées.

Les variables qui, en recherche, seraient considérées comme des artéfacts suscitent au contraire l'intérêt de l'évaluateur.

La question de la validité de la mesure en évaluation est liée au fait que la mesure en sciences sociales est, à très peu d'exceptions près, indirecte, car la plupart des concepts auxquels elle s'intéresse sont abstraits ou sont des construits théoriques. Ce qui signifie que les chercheurs en sciences sociales ne peuvent jamais être tout à fait certains que ce qu'ils mesurent *de facto* est bien ce qu'ils désirent mesurer (Nachemias et Nachemias, 1982). Cependant, l'incertitude concernant la validité de construit (variable théorique vague ou confuse, ou manque de congruence entre la variable théorique et la variable observable) ne pose pas un problème en évaluation, puisque la définition opérationnelle et les questions utilisées proviennent de ce qui se passe sur le terrain, et non d'une théorie. En évaluation, la validité de construit d'un instrument de mesure correspond à la congruence entre l'observation et l'étiquette de la variable étudiée (Peres et Yatsiv, 1994).

Durant les vingt dernières années, les auteurs ont plutôt penché vers une définition vague de la validité, permettant ainsi les jugements subjectifs. Par exemple, Cronbach (1983) affirme que la validité est plus subjective qu'objective. Krathwohl (1982) compare la validité à la persuasion, au sérieux et au consensus. Campbell, à qui revient la paternité des concepts de validité interne et externe, a pris presque entièrement ses distances par rapport à ces concepts (Campbell, 1986). Il suggère de formuler de nouvelles définitions, la validité interne devenant la validité locale, molaire (pragmatique et athéorique) et causale. Cette innovation met en doute l'idée qu'un nombre déterminé de traitements occasionne un réel changement dans une activité, un lieu et un temps spécifiques. Selon Campbell, on retrouve sur le terrain des relations de cause à effet dont aucune théorie ne peut expliquer l'origine et qu'on ne sait comment généraliser. Pour y arriver, on doit renverser l'ordre habituel des interventions, qui est d'aborder la théorie avant l'activité. La cause et l'effet que l'on constate sur le terrain et qui ne proviennent pas d'une théorie reconnue peuvent devenir un tremplin pour une nouvelle théorie.

La validité externe concerne la généralisation des résultats et les liens irréfutables pouvant être prédits entre les phénomènes empiriques. Puisque la recherche vise habituellement à généraliser des résultats obtenus pour pouvoir les appliquer à d'autres réalités (organismes, populations, situations,

temps) afin de formuler une loi universelle, une question typique de la recherche pourrait être : « Le développement de relations sociales entre des personnes vivant ensemble 24 heures par jour durant une période de temps prolongée peut-il être appliqué à d'autres personnes dans d'autres situations ? » Mais dans une évaluation, la consolidation d'un groupe de candidats à un cours pour pilotes de chasse, par exemple, peut très bien être la question à étudier et qui doit permettre aux commandants de prendre des décisions. Les facteurs d'évaluation auraient-ils la même importance pour des personnes qui dirigent un camp d'été pour jeunes et qui cherchent à augmenter le degré de satisfaction des participants ? Les évaluateurs en charge de l'enquête auprès des pilotes de chasse ne cherchent pas vraiment à savoir si les mêmes liens se forment entre les habitants d'une maison de retraite et ceux d'un pensionnat, alors que cette question peut très bien intéresser des chercheurs. La question de la validité externe est donc extrêmement importante pour la recherche, mais elle l'est moins pour l'évaluation.

D'un autre côté, la possibilité de généraliser l'effet d'un projet est tout de même une question importante. Des gestionnaires ou des entrepreneurs mettent toujours un projet en œuvre dans le but d'obtenir un effet désiré. L'évaluation peut vérifier si les effets du projet peuvent être généralisés ou si le projet a un plus grand potentiel d'application, si on s'intéresse à ces questions. En d'autres termes, on se questionne sur la possibilité de reprendre le traitement mis en œuvre par le projet, de l'appliquer à d'autres environnements, à d'autres populations. On peut donc généraliser des conclusions, mais non des résultats.

Lorsqu'on effectue, à partir d'un échantillon, une généralisation destinée à d'autres groupes de participants, il faut vérifier la représentativité réelle de cet échantillon. En évaluation, on s'intéresse à la généralisation de chaque partie du projet à d'autres et d'un ensemble de variables à un autre afin d'éviter de décrire le projet comme une série d'événements déconnectés les uns des autres. Par exemple, si la définition des tâches du personnel senior d'un projet se caractérise par des limites floues, il devient justifié de se demander si cela caractérise également les tâches de l'ensemble du personnel du projet. S'il y a un manque de familiarité avec les concepts centraux d'un projet dans une des écoles participant à ce projet (c'est-à-dire qu'il y a un problème dans une école en particulier), cette même lacune est-elle présente dans d'autres écoles du projet ? Les gestionnaires du projet ont-ils eu de la difficulté à transmettre leur message ?

Ces hypothèses quant à la possibilité de généraliser les résultats d'une évaluation aident l'évaluateur à dresser un portrait cohérent d'un programme, à repérer des phénomènes inhabituels (par exemple, une inefficacité dans une situation où l'on est généralement efficace) et à déterminer la source de ces problèmes (des gestionnaires inaptes à mobiliser les enseignants, lesquels ne leur font plus confiance). La possibilité de généraliser un résultat particulier procure aux évaluateurs un outil important pour comprendre le domaine dans lequel ils travaillent ; cela leur fournit des connaissances importantes et pertinentes pour eux-mêmes comme pour le projet dans lequel ils interviennent.

L'examen d'un nombre de variables significatives, dans le cadre d'une recherche structurée, peut faire partie d'un contexte plus vaste d'évaluation qui inclut d'autres examens, des variables additionnelles et des questions concernant les liens entre les événements ainsi que les effets de leurs interactions. Les concepts opérationnels de « variables isolées », d'« échantillons contrôlés » et de « généralisation » n'existent pas dans le domaine de l'évaluation. Dans la pratique, les variables ne sont pas isolées, les groupes examinés ne sont pas purs ou contrôlés, et la possibilité de généraliser est très limitée et même parfois nulle.

L'influence que peuvent avoir ces notions sur le succès ou l'échec d'un projet en particulier n'est généralement pas le genre de question que les chercheurs se posent, mais pour les évaluateurs, c'est l'essence même de leur travail. Les chercheurs choisissent leurs sujets de recherche en fonction de leur domaine de spécialisation ainsi que de l'intérêt et de l'importance que ces sujets présentent. Les évaluateurs choisissent les leurs en fonction de l'utilité probable des réponses qu'ils en tireront pour le projet sur lequel ils travaillent. En d'autres termes, c'est en fonction de la pertinence pour le projet qu'ils effectuent leurs choix.

L'importance de la pertinence

La pertinence d'une évaluation est extrêmement importante, puisqu'une part essentielle de sa valeur tient au fait qu'elle doive être en mesure de procurer une rétroaction efficace aux décideurs d'un projet. L'évaluateur fournit un service qui se doit d'être efficace, et l'efficacité et la pertinence de la rétroaction concernant le résultat d'un projet à un moment précis et impliquant une clientèle spécifique dépendent de la qualité de l'examen

effectué. En recherche, parce que le but est de pouvoir généraliser les résultats, on ne questionne pas la pertinence de ces résultats pour le domaine dans lequel la recherche est faite (à moins qu'on ne travaille dans le département de « recherche et développement » d'une entreprise, qui doit répondre à des besoins bien précis). Par ailleurs, il peut y avoir négociation entre l'évaluateur et les directeurs du projet ou les bailleurs de fonds, concernant les sujets qui sont pertinents ou non et l'allocation des ressources destinées à l'évaluation.

L'intérêt d'une évaluation est en soi une raison suffisante pour la justifier, mais cet intérêt ne suffit pas à formuler les questions de l'évaluation. Inversement, les sujets de recherche scientifique ne font pas l'objet de négociations, mais bien de discussions entre un chercheur et ses collègues. Les décisions concernant la manière d'examiner un sujet de recherche restent à la discrétion de ses initiateurs.

La mise en œuvre d'une procédure d'examen pour un suivi de variables généralisées et abstraites n'est pas vraiment utile pour les évaluateurs et leurs clients à la lumière du critère de pertinence. Et cela parce que les réponses qui peuvent en découler sont loin de la réalité des clients et de leurs préoccupations immédiates.

Pour être pertinente et utile, la rétroaction doit répondre aux conditions suivantes :

L'immédiateté. La rétroaction devrait être immédiate ou offerte dans un délai permettant d'effectuer une analyse des résultats et d'en tirer des conclusions qui contribueront à un changement ou à la planification d'un changement. Si la rétroaction est retardée trop longtemps, elle risque d'être inutile aux personnes à qui elle est destinée. Même dans le cas d'une évaluation sommative *post facto*, la rétroaction doit être disponible rapidement pour des fins de planification future.

La spécificité. La rétroaction doit être spécifique afin de donner des réponses en lien direct avec le projet faisant l'objet de l'évaluation. Ce qui signifie qu'elle s'adresse à la clientèle cible du projet, à ses gestionnaires et à ses planificateurs; elle ne devrait donc pas traiter de concepts généralisés.

Nous trouvons un bon exemple de cela dans un projet de renforcement de l'autonomie des femmes chez les Bédouins. La frustration de ces femmes peut devenir plus claire, pour les personnes qui dirigent ce projet, si on examine des facteurs concrets de leur situation. Par exemple, on ne

devrait pas dire, au cours de la rétroaction : « La structure patriarcale de cette société est menacée, aussi les hommes résistent-ils au projet parce que les femmes reçoivent un traitement différent. Il faut faire quelque chose pour réduire la perception de menace et la résistance qui en résulte. » Il serait préférable de se limiter à indiquer aux intervenants du projet que les hommes ne coopèrent pas, qu'ils ne permettent pas à leurs femmes de participer à certaines activités, et qu'il serait bon de rassembler les hommes pour leur expliquer les objectifs du programme.

Les résultats exprimés comme des lois universelles ou des concepts théoriques abstraits sont généralement inutiles aux gestionnaires et aux décideurs œuvrant sur le terrain ; de plus, ils ne sont pas toujours bien compris par les non-initiés au domaine scientifique.

Le langage du projet. La rétroaction doit être formulée à l'aide de la terminologie propre au projet ou, à tout le moins, dans un langage qui soit compris par ses gestionnaires et ses initiateurs, même si leurs définitions et leur terminologie ne correspondent pas tout à fait au jargon professionnel des évaluateurs. Si, dans la terminologie d'un projet, on utilise le terme « conseillers », on ne doit pas le remplacer par « chefs de groupe », « agents de changement », « coordonnateurs » ou « animateurs », même si on croit que ces termes sont plus justes. Si un projet parle de créer du pouvoir ou de l'autonomie dans une population ciblée, ce sont les concepts qu'on doit employer. On peut bien sûr négocier un changement en cas d'utilisation de concepts inadéquats, ou tenter d'introduire de nouveaux termes durant la mise en œuvre de l'évaluation, mais lorsqu'une terminologie particulière existe déjà, on se doit de l'utiliser. Cela est vrai pour différents types d'évaluation comme les évaluations formative, sommative, de processus et de résultats.

Les jugements de valeur. On ne peut parler de pertinence de la rétroaction sans aborder une des principales différences entre l'évaluation et la recherche, différence liée à la pertinence des résultats, soit l'aspect « jugement » inhérent à l'évaluation. L'histoire de la discipline de l'évaluation montre que le besoin du client est inscrit au cœur même de son développement, le rôle des évaluateurs étant d'interpréter et de juger les résultats des programmes. C'est aussi l'origine même de l'évaluation en tant que discipline. À la fin des années 1970, un comité fut en effet mis sur pied aux États-Unis avec pour mission de définir l'essence de l'évaluation en

tant que discipline. Ce comité a décidé d'insister sur l'importance de la décision que doit prendre l'évaluateur quant à la valeur de l'objet évalué (JCSEE, 1981). Cela vaut aussi pour les évaluations sommatives, qui se concentrent sur un seul résultat, soit mesurer l'atteinte des objectifs d'un projet, ce qui est une demande fréquente (Mor, 1992). Cela est particulièrement vrai en ce qui concerne les approches formatives si populaires de nos jours, qui discutent de la valeur des buts d'un projet, des dommages ou des pertes qu'il peut entraîner et de l'importance de ses effets.

La recherche scientifique tente d'être objective, et les résultats en constituent le message ultime. L'évaluation, parce qu'elle est un outil de prise de décisions, ne peut, et ne devrait d'ailleurs pas, éviter de faire des recommandations, même si celles-ci touchent des valeurs. Un processus d'évaluation qui n'impliquerait que la collecte de données, sans jugement porté sur leur essence, sans avis, sans proposition de solution de rechange et sans choisir entre elles ne serait ni possible ni souhaitable (Cordray, 1986; JCSEE, 1981; Levin-Rozalis, 1987, 1998; Lincoln et Guba, 1986; Scriven, 1967, 1983; Stake, 1969).

Le problème de la causalité

L'effort consenti pour comprendre la cause et la nature d'un phénomène, ainsi que la nature de la cause elle-même, représente un élément important de la rétroaction provenant d'une évaluation. On peut à tout le moins essayer de maîtriser les effets d'une cause lorsqu'on connaît celle-ci, mais les réponses concernant les liens entre variables ne sont généralement pas satisfaisantes. Nachemias et Nachemias (1982) soutiennent qu'il y a un dilemme fondamental entre la causalité et la généralisation : en effet, afin de s'assurer que l'évidence de la causalité est bien claire, on sacrifie souvent la possibilité de généraliser. Cela est bien sûr vrai pour la recherche scientifique. Mais en évaluation, tel que discuté plus haut, la possibilité de généraliser n'est habituellement pas aussi importante que la compréhension de la cause d'un phénomène.

Par exemple, des enfants participant à un projet (Levin-Rozalis et Bar-On, 1994) se plaignaient constamment de la nourriture, qu'ils qualifiaient de « nourriture d'école », alors qu'ils disaient vouloir de la « nourriture de maison ». On a tenté à plusieurs reprises d'améliorer la qualité des repas ou de préparer de la cuisine traditionnelle, mais en vain. Il n'y

avait pas de corrélation entre le type de nourriture et le nombre de plaintes. Une enquête approfondie a démontré que les enfants ne trouvaient pas vraiment que la nourriture était mauvaise, mais bien que celle-ci leur servait de moyen pour exprimer une insatisfaction face au projet, faute d'avoir d'autres moyens de le faire. Parce qu'ils considéraient comme un honneur le fait de faire partie de ce projet, les enfants n'osaient pas exprimer cette insatisfaction. En comprenant la nature du problème des enfants et donc la cause de leurs plaintes, les gestionnaires du projet ont pu parler ouvertement du problème et le régler. Après que ce problème fut résolu, les plaintes concernant la nourriture ont cessé.

Mor (1992) soutient que même une évaluation sommative, qui est menée alors qu'un projet est terminé, ne peut se satisfaire des seuls résultats du projet, mais doit examiner ce qui a été fait et comment, expliquer pourquoi les résultats sont ce qu'ils sont, définir les causes de cet état de fait. Cela ne signifie bien sûr pas qu'une explication causale existe nécessairement, qu'elle soit toujours facile à trouver, ni d'ailleurs que les évaluateurs doivent toujours déployer de grands efforts pour la découvrir.

3

La construction d'un jugement

Bernard Perret

Les définitions courantes de l'évaluation mettent l'accent sur l'idée de mesure : évaluer un programme ou une politique, c'est, pour l'essentiel, chercher à mesurer ses effets. Cette caractérisation a l'avantage d'être simple, mais elle reflète assez mal la réalité des pratiques. Le but de la discipline n'est pas seulement de mesurer les effets des actions publiques, mais également d'en comprendre les mécanismes. Par ailleurs, l'évaluation n'est pas une activité purement scientifique : c'est une démarche collective de *construction de jugements pratiques*, c'est-à-dire de jugements portés en vue de poser une action ou de prendre une décision. Or, tout jugement se cristallise dans un langage. Pour contribuer à l'articulation d'un jugement, les faits bruts doivent être transformés en *arguments*. De ce fait, l'évaluation revêt un caractère *rhétorique*, ce terme n'ayant ici aucune connotation péjorative, mais renvoyant seulement à l'idée d'une « pratique réglée » de l'argumentation. Il est suggestif, dans cette perspective, de penser le déroulement d'une évaluation comme celui d'un *procès*, modèle type d'une discussion soumise à des règles formelles portant sur le statut, les critères de validité et le poids relatif des arguments.

En établissant ce cadre d'analyse, on ne perd pas de vue les aspects techniques de l'évaluation, mais on les replace dans leur contexte. La recherche de preuves valides, objet central de la discussion méthodologique, s'inscrit

dans la perspective plus large d'une réflexion sur les fondements d'un jugement pratique. Il ne s'agit pas de s'affranchir des impératifs de la méthode scientifique, mais au contraire d'étendre l'exigence de rigueur à l'ensemble du processus social d'évaluation, en amont et en aval du travail de recherche, depuis la formulation des questions d'évaluation jusqu'aux techniques de mise en forme et de diffusion des résultats.

Les limites de l'analyse causale

On emploie fréquemment l'expression « recherche évaluative » (*evaluation research*) pour désigner la partie la plus « scientifique » du travail d'évaluation, à savoir la mesure et l'analyse des effets propres d'une action (typiquement un « programme ») sur la société.

À ses débuts, l'évaluation de programmes cherchait à imiter les sciences dites « exactes », comme la biologie. L'objectif était de mesurer les effets des interventions sur la société avec la même précision scientifique que, par exemple, celle que l'on exige dans les tests d'efficacité d'un nouveau médicament. Aujourd'hui encore, la recherche évaluative s'entend d'abord comme une activité de modélisation et de mesure. Il s'agit, idéalement, d'établir et de quantifier un lien de causalité entre un « traitement » et un effet recherché (par exemple, entre une mesure d'aide à l'emploi et la diminution du chômage). En pratique, cependant, cette ambition se heurte à deux types d'obstacles. En premier lieu, l'évaluation porte souvent sur des politiques complexes mêlant différents types de mesures (subventions, réglementations...) dont les objectifs sont reformulés au fil du temps et le contenu concret modifié au gré des situations locales. On peut illustrer cela par l'exemple des programmes de développement social des zones urbaines défavorisées, sans cesse remaniées et complexifiées au cours des années. Dans un tel cas de figure, l'analyse causale semble hors de portée, voire dénuée de sens. En second lieu, lorsque la nature des questions posées la rend envisageable, l'analyse causale bute souvent sur des obstacles techniques qui limitent sa portée. Mesurer un effet, c'est toujours comparer deux situations : d'une manière ou d'une autre, il faut observer ou reconstituer ce qui se serait passé si le programme n'avait pas été mis en œuvre. Il n'y a pas lieu de présenter ici en détail les méthodes utilisées à cette fin (expérimentation avec groupe de contrôle, protocole quasi expérimental avec estimation d'un modèle, etc.), mais simplement

de souligner que les conditions pratiques de leur mise en œuvre sont rarement réunies, tant elles supposent de protocoles d'observation rigoureux et de données complètes sur les phénomènes analysés.

L'évaluation s'intéresse autant à la compréhension qu'à la mesure

À supposer même que l'analyse causale débouche sur des conclusions précises, il est rare que l'évaluation puisse s'en contenter. Tout d'abord, même lorsqu'on parvient à mesurer un effet, il s'avère presque aussi important de comprendre comment et pourquoi il se produit. Chen (1990) prend l'exemple des médicaments : « Si une évaluation "boîte noire" montre qu'un nouveau médicament est capable de soigner une maladie mais ne fournit pas d'information sur ses mécanismes d'efficacité, les médecins hésiteront à prescrire ce nouveau médicament, parce que les conditions dans lesquelles il agit et les risques d'effets indésirables ne seront pas connus » (p. 18 et 42 ; voir aussi Chen, 2005).

L'analyse des mécanismes – qui repose sur des recherches qualitatives (monographies, enquêtes par entretien...), mais également sur l'activité auto-interprétative des acteurs eux-mêmes – vise notamment à déterminer à quelles conditions et dans quel contexte un programme est susceptible de produire les effets attendus. L'adjectif « qualitatif », qui s'oppose à « quantitatif », signifie simplement que l'analyse vise à décrire et à comprendre par des mots plus que par des chiffres. Un courant important en évaluation contemporaine, l'évaluation fondée sur une théorie (*Theory based evaluation*), recommande de centrer l'évaluation sur les mécanismes d'action ou, plus largement, la « théorie du programme ». La théorie d'un programme, c'est la structure logique des idées et des raisonnements sur lesquels il repose. Cette théorie peut se décomposer en segments élémentaires constitués, par exemple, de relations supposées entre des mesures d'incitation et le comportement des personnes ou des collectivités qui en bénéficient. Les évaluateurs en donnent parfois une représentation graphique (comme dans la méthode dite du « cadre logique »).

Concevoir une politique, c'est comme bâtir une théorie : cela revient à émettre des hypothèses sur les conséquences de certaines actions. Ces hypothèses sont souvent fragiles, fondées sur des idées reçues. C'est l'objet même de l'évaluation que de les soumettre à l'épreuve des faits. En pratique, la mesure des effets étant difficile, il suffit souvent d'observer ce qui

se passe, de décrire et d'interpréter le comportement des acteurs sociaux concernés, pour se faire une idée de la validité des raisonnements qui sous-tendent une politique.

L'évaluation doit viser une remise en question des idées des décideurs

Le jugement évaluatif n'est jamais le résultat mécanique de l'application d'une méthode – qu'elle soit quantitative ou qualitative. Comme tout jugement, c'est aussi un travail sur les représentations de la réalité, qui passe par la médiation de l'échange verbal. Comme l'indiquent Thomas Schwandt et Jennifer Greene (2006), « l'intérêt pour les techniques ou les méthodes ne doit pas obscurcir la nature foncièrement discursive de l'évaluation », le terme « discursif » signifiant ici « le fait que le vocabulaire et les concepts utilisés pour connaître et exprimer la valeur d'une politique sont construits socialement par des êtres humains ».

Il est important de comprendre que cette « construction sociale de la réalité » (Berger et Lupman, 1966) ne peut être circonscrite à l'évaluation. Il est rare que les méthodes de la recherche évaluative permettent de répondre à toutes les questions qui se posent au sujet d'une action publique. Les décideurs et les acteurs disposent d'ailleurs toujours d'autres *informations*, d'origines et de natures variées, qui leur permettent *a minima* de se forger une opinion : statistiques, articles de presse, points de vue exprimés par les groupes d'intérêt, etc. L'influence de ces informations n'est évidemment pas proportionnelle à leur pertinence scientifique. Certains faits impressionnent parce qu'ils frappent directement la sensibilité, créant ou renforçant des opinions qu'une analyse rigoureuse de la réalité aura bien du mal à déconstruire. Il en va ainsi, par exemple, des « faits divers », lorsqu'ils sont présentés par les médias comme des images emblématiques d'une société rongée par l'insécurité.

Toutes ces informations sont organisées, interprétées et pondérées par leurs récepteurs en fonction de leur vécu, de leurs inclinations idéologiques et de leurs besoins pratiques. Dès avant son lancement, les commanditaires d'une évaluation ont leurs idées sur la politique évaluée, sur ses résultats probables et sur ce qui, le cas échéant, « fait problème ». En d'autres termes, ils ont leur propre théorie sur cette politique.

Il est important que l'évaluation prenne en compte le point de vue initial des décideurs et des autres protagonistes de l'évaluation, non bien

sûr comme des vérités définitives, mais comme des hypothèses à vérifier et à relativiser ou à approfondir. Pour que l'impact d'une évaluation sur les représentations communes de la réalité soit plus important, celle-ci doit s'efforcer de remettre ces représentations en question, de les soumettre à des tests susceptibles de les invalider. Or, cela ne peut être fait qu'en se plaçant dans le même espace sémantique. Si les résultats des évaluations sont fréquemment ignorés ou négligés, c'est parce qu'ils sont souvent exprimés dans un langage que leurs destinataires ne reconnaissent pas, car trop différent de celui dans lequel ils formulent leurs préoccupations, leurs convictions ou leurs expériences. Le travail autour de la formulation des questions d'évaluation revêt de ce fait une importance cruciale.

Produire des données significatives au regard des représentations communes

Le but de l'évaluation n'est pas de produire des données scientifiques, mais de réduire l'incertitude et de renforcer la cohérence des décisions et des pratiques. En l'absence de preuves formelles de l'efficacité d'une action, on se satisfait souvent de «quasi-preuves», voire d'indices. Dans la vie réelle, on s'autorise souvent à tirer des conclusions générales à partir d'un petit nombre d'observations, pourvu qu'elles aient du sens, c'est-à-dire qu'elles s'intègrent dans un schéma rationnel et plausible préalablement construit (l'équivalent, pourrait-on dire, du «mobile» d'un crime). Nous sommes satisfaits quand nous disposons d'un faisceau d'indices convergents susceptibles d'emporter l'«intime conviction» d'une personne de bonne foi (c'est à dessein que j'emploie ici le vocabulaire du procès criminel). Il en va de même en évaluation. La phrase suivante, tirée d'un rapport d'évaluation, illustre la manière dont une observation portant sur un échantillon limité vient renforcer un point de vue préalablement constitué: «Il a été prouvé que les pathologies psychiatriques atteignant les parents pouvaient engendrer un risque majeur de mauvais traitement. Ainsi, dans notre série de 39 couples ayant exercé des mauvais traitements à l'égard de leurs enfants, 20 présentaient des troubles ou des antécédents psychiatriques importants.»

Trop d'information tue l'information

Les jugements pratiques sur l'action publique s'appuient donc sur des données nombreuses et variées. On pourrait en déduire, un peu hâtivement, que l'évaluation doit s'efforcer avant tout de rassembler le maximum d'information sur son objet, en partant du principe que l'on n'est jamais trop bien informé. Mais ce serait oublier qu'une surabondance de renseignements inadaptés est presque aussi nuisible qu'un manque d'information. Nous sommes tous encombrés de données non pertinentes – le phénomène des spams en fournit une bonne illustration – qu'il nous faut écarter au prix de dépenses non négligeables de temps et d'attention. Force est de reconnaître que toute information n'est pas bonne à prendre en considération. L'un de ceux qui ont le plus lucidement attiré l'attention sur ce phénomène est le prix Nobel d'économie Herbert Simon : « Dans un monde où l'attention est une ressource majeure des plus rares, l'information peut être un luxe coûteux, car elle peut détourner notre attention de ce qui est important vers ce qui ne l'est pas. Nous ne pouvons nous permettre de traiter une information simplement parce qu'elle est là » (Leca, 1993). Il suffit de parcourir un rapport sur un sujet quelconque pour vérifier qu'il contient en général un grand nombre de données inutiles.

Pour être utile, l'information évaluative doit donc être construite et formatée en vue de besoins spécifiques. La liste des questions à aborder doit être établie en fonction des attentes exprimées par les décideurs et porteurs d'enjeux. De même, au stade de la rédaction d'un rapport, les résultats, conclusions et recommandations doivent être hiérarchisés en tenant compte de ces attentes. Un rapport doit attirer l'attention sur les données les plus importantes et les « mettre en scène » de manière adéquate.

Le cerveau humain ne fonctionne pas comme un ordinateur

Cette conception « constructiviste » de l'évaluation peut être étayée par des considérations tirées de travaux contemporains sur le fonctionnement du cerveau humain (voir Varela, 1999, et Kolb et Whishaw, 2002). Nous comprenons mieux, aujourd'hui, que celui-ci ne fonctionne pas comme un ordinateur, mais comme un système en équilibre dynamique doté

d'une grande capacité d'autorégulation. Contrairement à un ordinateur, qui traite systématiquement les données qui lui sont soumises par ses périphériques d'entrée, notre cerveau choisit « librement » de prendre ou non en compte les informations qu'il capte par les cinq sens. Il les intègre en fonction de son propre programme, lequel garde la trace de tout le vécu antérieur de la personne. En d'autres mots, le cerveau consacre l'essentiel de son énergie à répondre à des questions qu'il se pose lui-même, et cela, dans la logique de son propre mouvement. Nous ne sommes pas préoccupés d'abord par ce que nous voyons et entendons, mais par la manière dont ces informations nouvelles peuvent nous aider à résoudre nos problèmes, affecter nos croyances et, plus largement, faciliter ou compliquer la gestion de nos états mentaux. Le lecteur d'un rapport d'évaluation, comme n'importe quel sujet pensant, utilise de manière sélective l'information mise à sa disposition en fonction de ses besoins et centres d'intérêt. L'évaluation lui sera d'autant plus profitable qu'elle parlera son langage et répondra à ses préoccupations.

L'approche cognitive des politiques publiques

L'évaluation, en résumé, est toujours un travail sur le langage et les représentations. Elle ne produit pas seulement des connaissances, mais aussi de nouvelles idées, une nouvelle manière d'appréhender les problèmes. Elle contribue à produire de nouveaux référentiels pour l'action publique. Cette conception de l'évaluation fait écho à « l'approche cognitive des politiques publiques ». Selon ce courant de pensée (voir Muller, 2006), il est erroné de considérer les politiques publiques comme le fruit de décisions rationnelles prises par des décideurs tout puissants et éclairés. Dans la réalité, l'action publique est la résultante instable d'une confrontation sans fin entre des forces sociales et des idéologies contradictoires. Et c'est de cet affrontement qu'émergent les conceptions qui servent de fondement intellectuel et idéologique à l'action publique. Comme l'indique Yves Surel (2000) : « Les politiques publiques sont le fruit d'interactions sociales qui donnent lieu à la production d'idées, de représentations et de valeurs communes. » L'évaluation peut être vue comme une méthode d'organisation de processus cognitifs et discursifs formalisés qui participent d'un processus plus large d'interactions sociales autour des enjeux de l'action publique. Il est important de préciser qu'elle assure cette fonc-

tion de production de référentiels à différents niveaux : celui de la conception des politiques, mais également celui de leur mise en œuvre par les acteurs de terrain. À chacun de ces niveaux, l'action publique est en effet partiellement redéfinie dans ses modalités et dans ses buts, et peut donc être remise en question par l'évaluation.

L'évaluation comme processus organisé de production d'idées partagées

Dans une opération d'évaluation, la connaissance n'est donc pas un but en soi ; il s'agit toujours de répondre à des questions que se posent les décideurs et acteurs de la politique. L'adaptation de la connaissance produite aux besoins de ses utilisateurs est révélatrice de la qualité du *processus* et de la *procédure* d'évaluation. Le *processus* est l'ensemble des interactions sociales suscitées par l'évaluation pour négocier son objet, ses questions, ses méthodes (le « projet d'évaluation ») et la formulation de ses conclusions. La *procédure* est la forme institutionnelle de ce processus. Elle comporte notamment la structuration d'une « scène évaluative » – distribution des rôles entre le commanditaire, le chef de projet, le comité d'évaluation, les experts et chargés d'étude, etc. La procédure recouvre également les normes méthodologiques et déontologiques qui règlent le processus – normes de qualité élaborées par les sociétés professionnelles d'évaluation ou normes internes d'une organisation.

Le choix des critères de résultats

En amont de la recherche évaluative, le processus d'évaluation commence par la formulation d'un questionnement évaluatif, lui-même appuyé sur une explicitation de la théorie du programme centrée sur l'articulation et la hiérarchisation de ses objectifs. En effet, les critères de résultats finalement sélectionnés dans une évaluation ne représentent qu'un sous-ensemble du vaste champ des phénomènes susceptibles d'être affectés par le programme. Avant même d'essayer de mesurer des résultats, on doit construire un cadre d'analyse et déterminer *a priori* les effets observables les plus à même de renseigner sur la réussite du programme, en regard de ses objectifs politiques (le « référentiel d'évaluation »).

Le choix de cadres de description

Toute opération de collecte d'information suppose la construction préalable de catégories, de nomenclatures, d'indicateurs, etc. Il est également nécessaire de procéder à des découpages temporels, spatiaux, etc. « Toute quantification suppose une mise en forme préalable de la réalité : il faut choisir les phénomènes à mesurer parmi une multitude de possibilités, définir des « unités statistiques » (donner un sens précis à des termes tels que ménage, entreprise, agglomération, quartier), élaborer des typologies et des nomenclatures, etc. Les chiffres n'ont de sens qu'à l'intérieur d'un cadre conceptuel sous-tendu par une représentation plus ou moins conventionnelle de la réalité. Celle-ci est toujours sujette à controverse, comme en témoignent les débats récurrents sur la mesure d'un « taux de chômage » ou d'un « taux de délinquance », par exemple (Perret et Seibel, 1993).

Une discussion pluraliste pour interpréter les résultats

En aval de la mesure, l'interprétation des chiffres repose toujours sur des modes d'appréhension de la réalité extérieurs aux données elles-mêmes. Il faut, d'une part, choisir parmi les résultats ceux que l'on considère comme les plus importants et, d'autre part, passer des résultats aux conclusions, aux jugements et aux recommandations. On a besoin pour cela de critères, de points de comparaison, qu'il faut aller chercher dans un stock de références préalablement constituées. L'exemple suivant, tiré d'une recherche évaluative, montre que cette activité d'interprétation est inhérente à la production de données chiffrées. La mesure d'un phénomène ne prend de sens que comparée à celle d'autres phénomènes supposément mieux connus du lecteur : « La fréquence des mauvais traitements infligés aux nouveau-nés est élevée (1,8 % de l'ensemble de la population). Pour donner une comparaison, cette fréquence de pathologie peut être rapprochée de celle des malformations fœtales dans l'ensemble de la population. »

Dans une recherche relevant d'une discipline scientifique (économie, sociologie, sciences de l'éducation, santé publique...), toute donnée est interprétée dans le cadre d'un schéma théorique propre à cette discipline. Le chercheur est censé maîtriser à la fois l'objet de la recherche, ses objectifs et son contexte. Il n'en va pas de même en évaluation. Dès lors qu'elle fait appel à des considérations politiques ou opérationnelles ne relevant pas d'un domaine scientifique spécialisé, l'interprétation des résultats

cesse d'être le monopole des experts pour devenir l'enjeu d'un débat pluraliste. L'organisation de ce débat est l'une des tâches principales des comités ou instances d'évaluation chargés d'en piloter les démarches.

L'apport de la théorie de l'argumentation

Une tradition qui remonte à Aristote

Les développements qui précèdent montrent l'intérêt d'analyser les pratiques évaluatives en tant que *pratiques argumentatives*. Pour Jürgen Habermas (1981, p. 41), une pratique argumentative peut être analysée selon trois points de vue : le processus (social), la procédure (les règles) et la production d'arguments. On retrouve ces trois aspects en évaluation. Il existe, dans le champ social, différents types de pratiques argumentatives, chacune étant caractérisée par la nature des arguments échangés (politiques, juridiques, moraux, esthétiques, etc.) et des « prétentions à la validité » qui y sont émises. La notion de « prétention à la validité » est importante ; elle signifie que tout argument prétend établir une vérité, mais que toutes les vérités ne se situent pas sur le même plan. Pour prendre un exemple classique, il n'est pas pertinent d'opposer un jugement esthétique ou moral à un argument de fait.

On peut faire remonter l'origine de la réflexion sur l'argumentation à la *Rhétorique* d'Aristote. Bien qu'il traite surtout de rhétorique au sens d'art de la persuasion (« la rhétorique est la faculté de considérer, pour chaque question, ce qui peut être propre à persuader »), Aristote observe qu'elle peut être vue dans une perspective plus large, comme un art de la discussion adapté à l'examen des questions incertaines : « L'objet de la rhétorique n'est pas tant de persuader que de voir l'état probable des choses par rapport à chaque question. » Aristote souligne fortement le lien entre la rhétorique et l'incertitude. Si la question de l'argumentation se pose, c'est qu'il y a incertitude, et donc doute ou pluralisme des opinions légitimes : « Nous délibérons sur des questions qui comportent des solutions diverses : car personne ne délibère sur des faits qui ne peuvent avoir été, être, ou devoir être autrement qu'ils ne sont présentés ; auquel cas il n'y a rien à faire qu'à reconnaître qu'ils sont ainsi¹. » Pour Aristote, l'une des

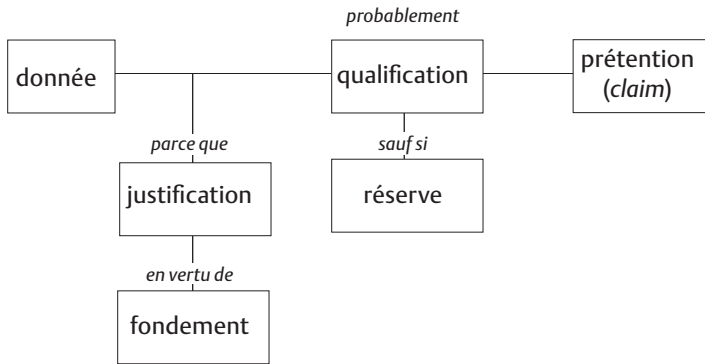
1. Aristote, *Rhétorique*, Paris, Livre de Poche, 1991, p. 81 et 87.

caractéristiques de la rhétorique est le recours fréquent à l'induction, c'est-à-dire aux exemples, faute de pouvoir démontrer de manière contraignante les thèses que l'on veut défendre.

Le schéma de Toulmin

Le logicien britannique Stephen Toulmin (1993) a posé les bases d'une réflexion moderne sur l'argumentation. Sa théorie de l'argument peut être résumée par un diagramme qui permet de visualiser la structure logique de toute forme d'argumentation :

FIGURE 3.1



- la *prétention* est le point (toujours sujet à disussion) que souhaite établir l'auteur de l'argument. Ce n'est pas une simple affirmation, mais une conclusion partielle qui justifie un effort d'argumentation spécifique ;
- la *donnée* est l'information nouvelle mobilisée comme support direct de la prétention ;
- la *justification* exprime le raisonnement utilisé pour établir un lien entre la donnée et la prétention ;
- le *fondement* est constitué par les faits ou raisonnements, supposés préalablement connus, qui viennent fonder, renforcer ou légitimer le principe contenu dans la justification ;
- la *qualification* indique le degré de certitude ou la force rationnelle que l'auteur attribue à la justification ;
- la *réserve* précise les circonstances ou les conditions qui peuvent réduire la force et la portée de l'argument.

Cette structure logique ne doit pas être confondue avec la structure grammaticale des phrases. Le schéma de Toulmin est rarement reconnaissable au premier coup d'œil. En effet, distinguer la donnée, la justification et le fondement n'est pas toujours évident. Prenons la phrase suivante : « La baisse récente du [taux de] chômage semble prouver l'efficacité des mesures gouvernementales. » Cet argument s'analyse comme suit :

- *prétention* : les mesures gouvernementales sont efficaces ;
- *donnée* : le taux de chômage baisse ;
- *justification* : la concomitance de la mise en œuvre des mesures et de la baisse du taux de chômage est l'indice d'un *rapport de cause à effet* ;
- *fondement* : on sait *par ailleurs* que les mesures gouvernementales sont, *par leur nature même*, susceptibles d'être efficaces ;
- *qualification* : elle est exprimée par un verbe de modalité, en l'occurrence « semble » ;
- *réserve* : aucune.

Toulmin insiste sur le fait que tous nos arguments ont une face cachée. Ils font appel de manière implicite à des éléments de contexte supposés connus et correctement interprétés. La validité d'un argument n'est donc pas une affaire de logique pure : elle dépend de stipulations supposées et acceptées par le public du débat. Dans l'exemple donné, il est clair que l'argument n'a guère de valeur si l'on ne croit pas au départ en l'efficacité probable des mesures gouvernementales. L'un des apports de Toulmin a été de souligner que l'on ne peut analyser les pratiques d'argumentation en se référant au modèle de la logique formelle. Le modèle pertinent serait plutôt celui de la procédure judiciaire, qui peut aider à prendre conscience de l'hétérogénéité des arguments mobilisés dans toute discussion orientée vers un jugement :

Les énoncés judiciaires ont plusieurs fonctions distinctes. Jugements sur une plainte, preuves d'identification, témoignages au sujet d'événements ou de controverses, interprétations d'un statut ou discussion de sa validité, exemptions de l'application d'une loi, plaidoyers en atténuation, verdicts, sentences : toutes ces différentes classes de proposition ont leur part à jouer dans la procédure judiciaire, et les différences entre elles sont loin d'être sans importance. Quand nous quittons le domaine particulier du droit pour considérer les arguments rationnels en général, nous sommes immédiatement confrontés à la question de savoir s'il ne faut pas les analyser à l'aide d'un ensemble de catégories d'une égale complexité. (Toulmin, 1993, p. 96)

Les types de données et de justifications employés en évaluation

Dans les évaluations, on trouve des arguments de nature *factuelle* (faits bruts, données chiffrées, constats et descriptions), mais également des arguments *logiques* (fondés sur un raisonnement logico-mathématique) ainsi que des arguments fondés sur des *témoignages*, des *points de vue* ou des *opinions*. Certains arguments ont un caractère *normatif*: ils s'appuient sur des valeurs ou des normes supposées reconnues. On trouve également des arguments d'*autorité* sous différentes formes: autorité d'un expert, d'une théorie, d'un corpus de connaissances, etc. Il est important de distinguer ces différents registres et de faire apparaître clairement leur articulation respective.

Tous les arguments n'ont pas la même portée. Les méthodes de la recherche quantitative ont pour but de produire des arguments quasiment contraignants reposant sur des certitudes de type mathématique. Il ressort cependant des travaux de Toulmin que le lien entre les faits établis et les conclusions que l'on en tire n'est jamais un lien strictement logique: il est assuré par des *justifications* elles-mêmes appuyées sur des *fondements* de natures diverses. Il en résulte que les arguments factuels ou scientifiques doivent toujours composer avec des arguments au statut plus incertain. Pour en rester au niveau le plus élémentaire, un chiffre doit toujours être *qualifié* pour permettre de porter un jugement sur une situation: l'augmentation du nombre de bénéficiaires d'une mesure peut être qualifiée de «rapide» ou «lente», l'évolution d'une situation peut être qualifiée de «satisfaisante» ou de «préoccupante», un effet peut être qualifié de «significatif», etc.

Les éléments d'une «bonne pratique» de l'argumentation

L'évaluation comme discussion méthodique

Par bien des aspects, l'évaluation s'apparente à une discussion ordinaire au cours de laquelle des arguments sont échangés dans le but de parvenir à un jugement mieux fondé et, autant que possible, partagé. Si le jugement évaluatif diffère du jugement ordinaire, c'est moins par son objet et la nature des informations mobilisées que par l'effort pour inscrire le processus argumentatif dans un cadre méthodologique et procédural rigoureux.

L'extrait suivant, tiré d'un guide méthodologique français, résume parfaitement ce point :

Contrairement à une recherche scientifique menée dans un cadre disciplinaire déterminé, qui privilégie un nombre limité de raisonnements, l'évaluation cherche à utiliser toutes les heuristiques, et à s'adapter aux conditions réelles de la délibération et de la prise de décision. Comme toute discussion ou réflexion menée à des fins pratiques, l'évaluation n'exclut *a priori* aucun élément d'information concernant son objet, quelle que soit sa nature (quantitative ou qualitative) et son origine, dès lors qu'il semble pertinent. [Mais] cet éclectisme n'est pas synonyme de confusion : contrairement à une discussion ordinaire, menée sans méthode formelle, l'évaluation s'efforce de ne pas mélanger les différents types d'arguments, mais de les hiérarchiser, de les pondérer et de relier chacun à des conclusions spécifiques. (Conseil scientifique de l'évaluation, 1996, p. 49)

« Discutabilité » versus persuasion

L'analogie du procès employée plus haut est suggestive, mais pas totalement pertinente. Une évaluation n'est pas un affrontement réglé de deux thèses contradictoires, mais un processus coopératif par lequel on cherche à établir un jugement raisonnable en regard de critères socialement légitimes (les objectifs d'une politique particulière ou, plus largement, l'intérêt général). Un rapport d'évaluation doit donc assurer la *discutabilité* des prétentions qu'il contient. Tout doit être fait pour qu'un lecteur quelconque puisse se faire une idée précise de la validité des arguments : connaissance des sources d'information et de leurs faiblesses éventuelles, limites inhérentes aux méthodes employées, etc. En d'autres termes, il convient de garantir la traçabilité des arguments.

L'explicitation du questionnement comme élément de la traçabilité des arguments

Un argument constitue toujours une réponse à une question. Le langage courant en porte la marque, puisque le terme « question » y est souvent associé à l'idée de signification – même en l'absence de point d'interrogation, par exemple quand on cherche « ce dont il est question » dans un texte. Dans la discussion courante, les questions restent souvent implicites.

En revanche, dans l'argumentation qui fonde une décision de justice pénale, les questions sont explicitement posées. Il en va de même en évaluation : la formulation dans un langage clair des questions auxquelles on veut répondre est un facteur de clarification de l'argumentation évaluative. Le rappel de la liste des questions examinées permet de déterminer plus clairement les prétentions et leur nature (ce que l'on veut démontrer et ce sur quoi on s'appuie dans ce but).

Les conséquences pratiques pour les démarches d'évaluation

La conduite d'un projet d'évaluation doit être orientée vers la construction d'un jugement raisonnable, « appropriable » et « discutable ». Cela exige notamment de :

- déterminer *ex ante* – *au préalable* – les enjeux cruciaux du point de vue d'un jugement orienté vers l'action, c'est-à-dire les sujets sur lesquels il semble important de lever une incertitude, ainsi que les idées qu'il importe de valider ou de préciser ;
- formuler les questions d'évaluation en des termes qui préfigurent les points cruciaux de l'argumentation que l'on cherche à bâtir ;
- affecter les moyens dont on dispose aux investigations les plus susceptibles de permettre de répondre aux questions les plus importantes du point de vue de leur impact sur les jugements pratiques des décideurs et acteurs clés de la politique ;
- être particulièrement attentifs, au moment de la rédaction du rapport d'évaluation, à la rigueur et à la clarté de l'argumentation. Le rapport d'évaluation doit expliciter l'objet, les objectifs et les méthodes mises en œuvre, sans oublier de préciser leurs limites. Il doit permettre au lecteur de faire la part entre les faits établis et les opinions majoritaires, les conclusions qui en sont déduites et les recommandations qui en sont tirées.

Appliquer l'exigence de rigueur aux processus sociocognitifs

Notre insistance sur les dimensions discursive et interprétative de l'évaluation ne doit pas s'entendre comme une relativisation des enjeux du travail plus strictement technique de mesure des effets d'une action. Il

s'agissait plutôt d'attirer l'attention sur le fait que la phase de collecte et de traitement de l'information est toujours inscrite dans un processus plus large d'interactions sociales diversifiées dont la bonne organisation fait partie du travail d'évaluation. Ces interactions participent pleinement de l'efficacité de l'évaluation, non seulement sur les plans social, managérial et politique, mais également sur le plan strictement cognitif. La connaissance sociale n'a en effet de valeur, d'utilité et même de consistance qu'en tant qu'elle est reliée à des questions et à des jugements. Or, l'élaboration de ceux-ci relève d'une forme de travail intellectuel collectif que l'on peut rendre plus efficace et pertinent en le soumettant à des règles méthodologiques simples.

De ce point de vue, la réflexion sur les méthodes et les démarches d'évaluation a une grande portée épistémologique et politique trop peu soulignée à notre sens. Elle suggère la possibilité et l'intérêt d'appliquer les exigences de la rigueur et de l'objectivité qui ont fait le succès de la science à une classe plus étendue de processus sociocognitifs. On entrevoit ainsi la possibilité de développer une méthode et des pratiques visant à rendre plus réfléchi, organisé et systématique le travail sur soi des sociétés humaines qui, pour l'heure, apprennent peu et lentement de leurs expériences.

Pour aller plus loin :

Deux ouvrages classiques, déjà anciens, sur la théorie de l'argumentation :

PERELMAN, C. et L. OLBRECHTS-TYTECA (1958), *La nouvelle rhétorique. Traité de l'argumentation*, Paris, Presses universitaires de France.

TOULMIN, S. (1993), *Les usages de l'argumentation*, Paris, Presses universitaires de France.

Ouvrages abordant l'évaluation du point de vue de la discussion et de l'argumentation :

HOUSE, E. R. et K. R. HOWE (1999), *Values in Evaluation and Social Research*, Thousand Oaks, Sage Publications.

PERRET, B. (2001), *L'évaluation des politiques publiques*, Paris, La Découverte.

— (1995), « La construction du jugement », dans Conseil scientifique de l'évaluation, *L'évaluation en développement 1994*, Paris, La Documentation française.

SCHWANDT, T. A. (2005), « Politics of Evaluation », dans S. MATHISON (dir.), *Encyclopedia of Evaluation*, Thousand Oaks, Sage Publications.

Internet

PERRET, B. (1996), « Évaluation, gouvernance et procéduralisation de l'expertise », Intervention au séminaire sur « Les mutations contemporaines de l'art de gouverner » organisé par la Cellule de prospective de la Commission européenne, Bruxelles, <http://perso.orange.fr/bernard.perret/eva-proc.htm>

4

L'analyse des politiques publiques

Steve Jacob et Christine Rothmayr

L'évaluation de politiques ou de programmes publics est traditionnellement présentée comme une pratique ancrée théoriquement et méthodologiquement dans de nombreuses disciplines, dont l'économie, la psychologie et la sociologie. Son lieu de naissance précis reste toutefois plus difficile à déterminer, même si certains auteurs n'hésitent pas à affirmer que « la science politique est probablement le lieu de naissance de l'évaluation » (Schneider, 1986, p. 222 ; notre traduction). Dans ce chapitre, nous n'avons pas la prétention de trancher définitivement ce débat. Nous nous attarderons davantage à décrire la parenté entre l'évaluation et l'analyse des politiques publiques, un champ de la science politique actuellement en pleine expansion.

Plusieurs éléments rapprochent en effet l'évaluation et l'analyse des politiques publiques. *Primo*, il s'agit de processus centrés sur l'examen de l'intervention de l'État, et plus particulièrement des ministères et des organismes qui le composent, par ses actions concrètes et ses réalisations observables. Ces études appréhendent leur sujet par l'angle d'un objet politique commun qui, bien souvent, est le secteur de l'activité gouvernementale (par exemple, emploi, économie, environnement...). *Secundo*, l'évaluation et l'analyse des politiques alimentent le répertoire des

connaissances et des savoirs mobilisables par les décideurs et les gestionnaires. Il s'agit donc ici de la constitution d'un savoir davantage orienté vers la résolution de problèmes que vers la production de savoirs fondamentaux. Les chercheurs et les évaluateurs développent un langage commun en vue de « dire la vérité aux décideurs¹ », ce qui les place parfois en position de conseiller du prince (Duran, 1997). *Tertio*, les travaux d'évaluation et d'analyse des politiques accordent une place importante au jeu des acteurs dans l'étude des processus, afin de prendre en considération les enjeux de pouvoir comme nous allons le faire dans ce chapitre. En règle générale, tous les acteurs qui se trouvent dans le périmètre de la politique ou du programme considéré font l'objet d'une attention particulière, qu'ils relèvent du secteur public (parlementaires, ministres, gestionnaires, etc.) ou du secteur privé (groupes d'intérêt, ONG, société civile, etc.).

Mais il existe également des éléments de dissemblance entre l'évaluation et l'analyse des politiques, comme l'atteste le scepticisme de Brewer et deLeon (1983) à l'égard des travaux évaluatifs, dont les standards théoriques sont élevés mais qui, selon ces auteurs, se concrétisent difficilement. La principale différence réside dans le jugement de valeur porté par l'évaluateur sur la politique ou le programme qu'il étudie. À ce sujet, Vincent Lemieux (2006a) considère que « l'analyse des politiques sert à décrire ou à expliquer comment se réalisent les politiques dans leurs différentes phases, alors que l'évaluation de programmes vise plutôt à juger si les programmes sont conformes à des valeurs prescrites ou présumées ». Et ce jugement n'est pas sans conséquences sur les programmes, les décideurs ou les gestionnaires au sein des ministères et organismes. C'est à ce titre que « l'évaluation peut être considérée comme un processus politique, c'est-à-dire comme un ensemble de relations de pouvoir entre les acteurs concernés » (Lemieux, 2006e). En effet, l'évaluation, comme les autres étapes des politiques publiques que nous décrirons ici, ne peut être appréhendée sans prendre en considération le fait qu'elle est une activité politique qui s'appuie sur une démarche méthodologique rigoureuse. Pour parvenir à ses conclusions et à ses recommandations, et lorsque celles-ci sont retenues, l'évaluateur lance un processus d'apprentissage et de changement plus ou moins important du programme évalué. Nous reviendrons

1. Notre traduction du célèbre « *speaking truth to power* » de Wildavsky (1979).

un peu plus tard sur les différentes formes d'utilisation de la démarche évaluative, qui ne se cantonnent pas à la description de constats, mais qui ont par la suite auprès des décideurs et des gestionnaires des répercussions qui, dans certains cas, peuvent entraîner l'abolition du programme.

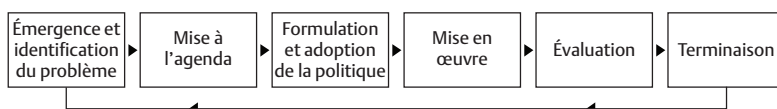
Afin d'établir des liens entre l'évaluation et l'analyse des politiques publiques, nous allons structurer notre explication autour du cycle de la politique. Les grilles d'analyse séquentielles qui découlent des travaux de Lasswell (1956) et de Jones (1970) permettent de comparer une politique ou un programme public à un organisme vivant dont la croissance est rythmée par des étapes successives allant de sa naissance jusqu'à sa mort ou sa renaissance. Une grille séquentielle regroupe donc, de façon ordonnée, les étapes qui structurent le développement d'une politique ou d'un programme public. Les différentes étapes du cycle de la politique varient selon les auteurs. Dans ce chapitre, nous nous concentrerons sur cinq étapes :

1. La prise en compte de l'émergence et de l'identification d'un problème permet de comprendre comment un problème privé devient un problème social, et donc d'appréhender les origines d'une éventuelle intervention publique par l'étude des controverses.
2. La mise à l'agenda est l'étape au cours de laquelle un décideur (parlementaire, élu, gestionnaire), qualifié d'entrepreneur politique, s'implique activement dans la promotion des intérêts qui se sont construits précédemment, et cela, en vue de convaincre de la pertinence politique de l'action de l'État en la matière. Cette implication résulte bien souvent de la mobilisation d'acteurs, de ressources et de stratégies variées (Cobb et Rochefort, 1993).
3. La formulation et l'adoption d'une politique ou d'un programme impliquent l'opérationnalisation et l'appréciation de différentes possibilités envisageables pour résoudre le problème à l'origine d'une intervention publique.
4. La mise en œuvre a trait à l'application des décisions adoptées précédemment. Au cours de cette étape, il arrive que des adaptations et des ajustements soient nécessaires pour parvenir à concrétiser les intentions des décideurs (Pressman et Wildavsky, 1973).
5. L'évaluation et la terminaison s'interrogent sur l'évolution et la pérennité des politiques et des programmes publics. L'évaluation n'étant pas une fin en soi, l'étude de la terminaison permet d'observer l'utilisation

de ses conclusions et recommandations quant à l'apprentissage, à la continuité ou à l'abolition du programme, et le suivi qui en est donné (deLeon, 1987).

FIGURE 4.1

Les étapes du cycle de la politique publique



Source : Inspiré de Muller, 2003, p. 24.

Cette schématisation du cycle d'une politique publique est une fiction, tant pour les analystes des politiques publiques que pour les évaluateurs. Les analystes savent que, bien souvent, ces séquences s'intercalent, se chevauchent, voire se superposent (Sabatier, 1997). De leur côté, les évaluateurs savent que l'évaluation n'intervient pas uniquement à la fin du cycle des politiques en vue de le boucler. L'évaluation vise à produire un savoir sur le fonctionnement et les effets d'une politique en l'appréciant à l'aune de plusieurs critères comme l'opportunité, la pertinence, l'efficacité et l'efficience. Ces connaissances servent à piloter la conduite de l'action publique.

De ce point de vue, et contrairement à la vision suggérée par l'analyse séquentielle des politiques, il est illusoire et artificiel de considérer que l'évaluation n'a lieu qu'après l'étape de la mise en œuvre. En tant qu'instrument de pilotage, il est possible de mobiliser l'évaluation au cours de chacune des phases d'une politique ou d'un programme. Une évaluation *ex ante* peut en effet alimenter les étapes de l'émergence d'un problème, de la mise à l'agenda ou de la formulation d'une politique, puisque, s'il s'agit par exemple de grands projets environnementaux ou d'infrastructure, ses résultats éclairent les décideurs sur leur faisabilité et les impacts qui en sont attendus. Au moment de la mise en œuvre d'une politique, les décideurs ou les gestionnaires publics peuvent commander une évaluation concomitante afin de documenter le processus de la mise en œuvre et d'y apporter des adaptations. Enfin, après quelques années de mise en œuvre d'une politique, certains souhaitent identifier, mesurer et analyser les effets de l'action publique au moyen d'une évaluation *ex post* (il s'agit du

type d'évaluation dont il est le plus souvent fait mention dans la littérature).

La littérature en administration publique nous enseigne que l'évaluation n'est qu'un instrument de pilotage parmi d'autres (par exemple, vérification, suivi, audit...) et qu'il n'est pas toujours évident de les distinguer avec précision. Ainsi, avec la montée en puissance du paradigme managérial, centré sur la performance, l'évaluation apparaît davantage comme un instrument de gestion qu'un outil de contrôle (Segsworth, 2004). Néanmoins, le principal élément qui permet de distinguer l'évaluation des autres instruments de pilotage de l'action publique réside dans l'attention qu'elle porte aux effets d'une politique ou d'un programme, c'est-à-dire aux changements (positifs ou négatifs, souhaités ou inattendus...) engendrés auprès des acteurs. De ce point de vue, les autres instruments de pilotage ne disposent pas des outils méthodologiques leur permettant d'identifier et d'isoler avec précision les effets imputables à l'intervention publique. Enfin, l'évaluation peut entraîner un questionnement et donc une remise en cause de la pertinence ou de l'opportunité des décisions préalables.

Cependant, malgré les limites traditionnellement énoncées à l'encontre de cette approche séquentielle, il nous apparaît tout de même utile de rattacher cette schématisation au processus d'évaluation. En effet, il nous semble opportun d'utiliser ce cycle pour analyser et comprendre la démarche évaluative. Cette déconstruction analytique permet d'isoler les moments clés du processus et de se concentrer sur les interactions entre les différents acteurs de l'évaluation (commanditaires, évaluateur, évalués). Au-delà de cette schématisation, nous nous attarderons également sur les principales théories d'analyse des politiques publiques qui nous aident à mieux comprendre le déroulement du processus évaluatif. À cet effet, la littérature sur l'analyse des politiques enrichit la compréhension de la pratique évaluative. Dans ce chapitre, nous allons nous appuyer sur l'analyse des politiques publiques pour mieux définir la pratique de l'évaluation et présenter les principaux enseignements théoriques relatifs à chacune des phases définies ci-dessus. Cela nous permettra de mettre en évidence les facteurs clés que les évaluateurs doivent avoir à l'esprit pour surmonter les écueils qui jalonnent le processus évaluatif.

La justification de l'évaluation : émergence et identification du problème

Dans une dimension centrée sur la résolution des problèmes, l'émergence d'une politique ou d'un programme résulte d'une volonté de changement en vue d'améliorer une situation vécue par certains membres de la collectivité (par exemple, chômage, délinquance, pollution...). Pour cela, il est indispensable que la situation problématique à laquelle font face les bénéficiaires pressentis d'une intervention publique soit érigée en problème collectif ou public, c'est-à-dire que celle-ci soit sortie de la sphère confidentielle ou privée dans laquelle elle se trouve à l'origine (Garraud, 1990). L'évaluation est parfois présente dans cette vision « traditionnelle » de l'émergence d'une politique, puisqu'il lui arrive d'alimenter cette première étape du cycle de la politique. Dans certains cas, l'évaluation attire l'attention du public sur un problème spécifique et s'avère donc être une des sources potentielles d'information qui permettent d'identifier un problème autour duquel une demande, voire une revendication sociale, pourra éventuellement se structurer. En l'absence de cette mobilisation et de cette reconnaissance sociale de la situation privée problématique, celle-ci ne sera pas thématisée politiquement et ne donnera pas lieu à une intervention publique (Knoepfel *et al.*, 2001). L'intervention publique résulterait donc d'un processus structuré en différentes étapes : l'identification d'un problème (*naming*), la désignation des causes et des responsabilités (*blaming*) et la revendication d'un changement (*claiming*) (Cobb et Ross, 1997). En prenant en considération tous ces éléments, il est entendu que la façon dont le débat va s'organiser influencera considérablement la suite des opérations. De ce point de vue, les analystes de politiques observent couramment que la structuration du débat oriente le registre des solutions dans lequel puiseront ensuite les décideurs.

La réalisation d'une évaluation n'est possible que si des acteurs en manifestent le désir. La nécessité, et parfois même l'urgence, d'une évaluation résulte de l'apparition de problèmes sérieux dans le fonctionnement d'un programme ou de déficits de mise en œuvre dans une politique, qui s'observent sous la forme de dysfonctionnements de la gestion publique. Dans ce cas, la crise devient un révélateur montrant l'urgence d'entreprendre une évaluation en vue d'alimenter le débat à l'aide de données probantes (*evidence-based policy*) ou de faire taire une polémique ou une controverse publiques. À cet égard, les décideurs ou gestionnaires

qui doivent affronter une crise insistent sur la nécessité de faire toute la lumière avec transparence sur ce phénomène (Jacob, 2006) ; l'évaluation externe constitue alors une voie commode, puisqu'elle permet de s'appuyer sur l'analyse d'une tierce partie (l'évaluateur) présentée comme neutre, experte et impartiale.

Dans d'autres cas, la réalisation d'une évaluation est prévue par le décideur lui-même, qui peut adopter une clause évaluative. Les clauses d'évaluation sont des dispositions que le législateur insère dans une loi et qui rendent obligatoire la réalisation d'une évaluation dans un délai plus ou moins imposé. Une telle clause peut être plus ou moins détaillée, allant de la simple intention d'évaluation à la mention des conditions pratiques de la démarche évaluative. Ces clauses permettent aux évaluateurs d'anticiper les attentes du commanditaire et favorisent un travail préparatoire en matière de collecte des données, d'élaboration d'instruments de mesure et de suivi, ainsi que de planification de l'exercice.

Le type de raisonnement a des conséquences sur les objectifs de l'évaluation (notamment sur l'implication, la collaboration, voire la conflictualité) à toutes les étapes ultérieures : au cours de la définition du mandat, au moment de l'implication des acteurs pendant la réalisation de l'évaluation et, bien sûr, lors de l'utilisation, au moment du suivi, des conclusions et des recommandations. Si l'évaluation est « imposée », elle apparaît comme un instrument de contrôle aux yeux de ceux qui sont présentés comme étant les responsables de la situation dénoncée initialement. Ils auront peur, se méfieront de l'exercice et tenteront d'en freiner la progression. On peut aussi dénoncer la partialité ou l'instrumentalisation politique des résultats produits. En conclusion, à l'issue de cette première étape, il est indispensable que l'évaluateur s'interroge sur l'origine de son mandat (d'où vient-il ? qui l'a défini ?) afin d'anticiper la suite du processus.

La commande de l'évaluation : mise à l'agenda

De nombreux auteurs considèrent que la phase de la mise à l'agenda est cruciale pour la formulation et la mise en œuvre ultérieures des politiques et des programmes publics (Kingdon, 1984 ; Baumgartner et Jones, 1993). Leurs explications s'appuient sur deux raisons principales. En premier lieu, la phase de mise à l'agenda apparaît comme un filtre qui permet de déterminer quels sont les enjeux qui figureront, ou non, sur l'agenda

politique. Les analystes de politiques ont démontré qu'il ne suffit pas de se demander quels sont les acteurs qui parviennent à faire émerger un enjeu sur l'agenda politique pour comprendre ce processus. En effet, les analyses s'enrichissent lorsque l'on prend en considération les acteurs qui ont le pouvoir d'entraver l'inscription d'un problème à l'agenda (Bachrach et Baratz, 1963). En second lieu, les analystes qui se concentrent sur cette étape insistent sur l'importance d'étudier la manière dont un problème ou un enjeu est (re)configuré au cours de cette étape du cycle de la politique. À ce stade, la définition et la traduction d'un problème dépendent des intérêts et des valeurs des acteurs impliqués, et se répercutent sur la définition des objectifs de la politique à venir et sur le type de solutions qui seront finalement retenues (Stone, 1989 ; Cobb et Rochefort, 1993).

Ces travaux peuvent alimenter les réflexions sur la décision d'entreprendre, ou non, une démarche évaluative. Dans cette perspective, la commande d'une évaluation ne se limite pas à ses aspects « technocratiques », mais revêt une dimension politique importante.

Tout d'abord, la décision d'évaluer un programme ou de se concentrer sur une de ses dimensions précises est déterminée par la fonction occupée par le commanditaire potentiel et par ses intérêts, mais aussi par ses connaissances sur le programme concerné. C'est à cette étape qu'on détermine si cette évaluation répond à un besoin général et qu'on en définit les objectifs spécifiques. Les réponses aux questions « Pourquoi entreprendre une évaluation ? » et « Quels sont les objectifs poursuivis par le commanditaire ? » dépendent fortement des acteurs qui amorcent l'exercice et de leur capacité d'inscrire la nécessité d'une évaluation à l'agenda administratif. Dans une perspective pluraliste, nous pouvons imaginer que le « jeu évaluatif » est ouvert à tout le monde (Dahl, 1961). Ce point de vue est d'ailleurs renforcé par les théoriciens de l'évaluation qui plaident pour la prise en compte d'un éventail d'intérêts élargi (Patton, 2000 ; Fetterman, 2001) et qui encouragent la démocratisation de la démarche évaluative (House et Howe, 2000). Cependant, les responsabilités et les prérogatives ne sont pas les mêmes pour toutes les parties prenantes en présence. La capacité pour un acteur d'entreprendre ou non une évaluation est influencée par la manière dont la pratique est institutionnalisée auprès des organes parlementaires, gouvernementaux ou administratifs (Jacob, 2005), ou par l'existence plus ou moins systématisée de clauses évaluatives.

Dans la pratique, la décision d'entreprendre une évaluation provient bien souvent des fonctionnaires responsables de la mise en œuvre de la politique ou du programme, des gestionnaires publics ou des institutions gouvernementales de vérification et de contrôle (Segsworth, 2002). Les fonctionnaires de première ligne peuvent également concourir à la prise de conscience du besoin d'évaluer une politique ou un programme, même si leur influence reste limitée, puisque leurs capacités décisionnelles sont peu développées. Lorsqu'un programme ou une politique est l'objet d'une controverse, il arrive que les décideurs (parlementaires, ministres, hauts fonctionnaires...), les groupes d'intérêt ou même les citoyens fassent pression pour conduire une évaluation (Jacob, 2007, p. 185-190). Toutefois, cela représente plutôt l'exception que la règle.

Par ailleurs, les motivations qui amènent les acteurs à entamer une évaluation ne sont pas toujours transparentes. Bien souvent, les raisons légitimes côtoient des motivations plus stratégiques, voire carrément intéressées. C'est également le cas dans le processus décisionnel classique, puisque, comme le démontre la littérature sur l'analyse des politiques, le fait de résoudre un problème n'est pas la seule explication à prendre en considération pour comprendre les justifications qui conduisent à l'élaboration d'une nouvelle politique. Dans le même ordre d'idées, en observant la pratique évaluative, nous constatons que l'amélioration d'un programme ou d'une politique, la mise en place d'un processus d'apprentissage organisationnel, la transmission d'informations aux décideurs ou le respect d'obligations évaluatives ne sont pas les seules motivations qui poussent les acteurs à entreprendre une telle démarche. Les intérêts cachés amplifient les risques d'instrumentalisation de l'évaluation. Il s'agit, par exemple, de l'ajournement d'une décision, de la légitimation d'une décision déjà prise, des tentatives d'apaiser les résistances au sein d'une administration, de l'amélioration de l'image d'une unité ou d'une organisation, et, par extension, du renforcement de sa position au sein de l'appareil administratif (Weiss, 1998).

Il est évidemment important qu'un évaluateur externe connaisse le contexte dans lequel a été prise la décision d'entamer une évaluation. En répondant aux questions « Comment est-on parvenu à cette décision ? » et « Y a-t-il eu des résistances, voire des oppositions, à la volonté de réaliser une évaluation ? », l'évaluateur disposera d'une information utile à la réalisation de son mandat. Il pourra adopter un comportement adéquat,

puisqu'il sera à même d'opérer des choix éclairés en matière de consultation, d'implication ou de collaboration entre les différentes parties prenantes. Les chercheurs qui s'intéressent davantage à la mise en œuvre des programmes ou des politiques ont d'ailleurs mis en évidence la nécessité de se référer à ces facteurs contextuels pour comprendre les positionnements ultérieurs des acteurs.

En répondant aux questions « Quels sont les acteurs qui ont défini le problème et proposé des solutions? » et « Quels sont les acteurs qui ont été impliqués ou exclus du processus? », il est possible d'anticiper la manière dont se déroulera la mise en œuvre de la politique et le comportement des acteurs associés ou non à cette étape. Il en va bien sûr de même en évaluation. C'est la raison pour laquelle l'évaluateur pourrait identifier et recenser systématiquement toutes les parties prenantes (*stakeholders*) directement ou indirectement concernées par la politique ou le programme à évaluer. À ce stade, il convient de dépasser les apparences et de sonder méticuleusement l'ensemble des intérêts qui seront en jeu au cours de l'évaluation. En effet, en identifiant les parties prenantes, leurs aspirations et leurs besoins, l'évaluateur peut pressentir les relations de pouvoir auxquelles il devra potentiellement faire face. Cela lui permettra d'anticiper les points de tension éventuels ou les moments où il aura possiblement à reformuler les attentes relatives à son travail. Ceux-ci peuvent d'ailleurs surgir dès les prémisses de l'évaluation, c'est-à-dire lors de la définition du mandat, comme nous allons le voir dès maintenant.

La définition du mandat évaluatif : formulation et adoption de la politique

Le processus décisionnel a depuis longtemps attiré l'attention de nombreux chercheurs en analyse des politiques. Plusieurs théories expliquant la manière dont les décideurs élaborent et fondent leurs choix ont été proposées. Il y a plus d'un demi-siècle, Herbert Simon, considérant que le modèle de l'acteur rationnel, l'*Homo economicus* dont les choix se fondent sur la maximisation de ses propres intérêts, ne permettait pas de comprendre le processus politique, a enrichi le cadre d'analyse en y intégrant le concept de rationalité limitée (Simon, 1945). Dans cette perspective, la rationalité parfaite n'existe pas, puisque les connaissances sont incomplètes et fragmentées, les acteurs n'anticipant pas toujours précisément les conséquences de leurs actions et n'étant pas aptes à analyser

complètement tous les aspects d'un problème. Dans ces conditions, leurs capacités de réflexion sont limitées; ils font donc face aux problèmes et prennent leurs décisions de manière fragmentée plutôt que simultanée. De plus, les routines et les autres solutions auxquelles ils ont déjà eu recours influencent leurs choix. En conséquence, les décisions qui sont adoptées permettent de satisfaire les objectifs poursuivis plutôt que de maximiser les gains. Par la suite, Lindblom a rompu encore plus radicalement avec la figure du décideur rationnel. De son point de vue, la décision politique se caractérise par l'incrémentalisme (Lindblom, 1979). Les décisions ne sont pas le résultat d'analyses planifiées ou de réflexions globales. Au contraire, les acteurs analysent les problèmes qu'ils rencontrent par rapport aux solutions disponibles; ils fondent leurs choix sur ce qui existe réellement et procèdent à de rapides comparaisons en vue de déterminer la pertinence de leur adoption. Ainsi, les politiques évoluent d'une manière progressive, comme le démontrent en effet plusieurs études accréditant le fait que les décisions politiques se fondent bien plus sur l'éventail des politiques existantes que sur l'innovation (Rose et Davies, 1993; Pierson, 2000).

Les concepts de culture ou de pratique évaluative attirent l'attention sur le fait que les valeurs, les routines et les procédures d'une unité, d'un département ou d'une organisation (ministère, agence...) peuvent elles aussi affecter sérieusement l'élaboration d'un mandat d'évaluation, bien souvent matérialisé sous la forme d'un cahier des charges.

En fait, en rédigeant un mandat d'évaluation, les décideurs publics reproduisent très souvent les habitudes et procédures politico-administratives ancrées dans leur environnement. Comme nous l'avons déjà mentionné, les caractéristiques des dispositifs d'institutionnalisation de la pratique évaluative influencent ce processus et sont cruciaux pour comprendre les éléments qui affectent la préparation d'un mandat et, ultérieurement, les résultats de l'exercice d'évaluation. Dans le même ordre d'idées, le profil de l'évaluateur aura également des répercussions sur l'ensemble de l'exercice. En effet, une évaluation réalisée par un organisme en charge d'une vérification (comptable et financière) et d'une évaluation n'aboutira pas nécessairement aux mêmes démarches, conclusions et recommandations qu'une évaluation confiée à une unité interne d'évaluation ou à un évaluateur externe (Segsworth, 2002, 2005; Saint-Martin, 2004).

Au cours de cette étape, il est nécessaire que le commanditaire définisse les objectifs de l'évaluation. La littérature nous enseigne que ces derniers sont nombreux, et qu'ils vont de la transmission d'informations aux décideurs jusqu'à la suppression d'un programme ou d'une politique, en passant par la mise en œuvre d'un processus d'apprentissage organisationnel au sein de l'administration en charge de la politique ou du programme évalué (Weiss, 1998). Les finalités de l'évaluation sont donc bien souvent étroitement liées aux ambitions poursuivies par celui qui va ultimement disposer de l'information produite par cet exercice. De plus, la réalisation d'une évaluation, en termes pratiques et concrets, requiert des décisions au sujet de l'approche privilégiée (par exemple, guidée par les théories participative, réaliste, émancipatrice...), du choix de l'évaluateur (en cas d'évaluation externe), mais aussi du devis évaluatif à appliquer, c'est-à-dire des méthodes retenues, de l'agenda et de la planification, du budget octroyé et de la valorisation souhaitée des résultats. Contrairement aux apparences, ces choix sont loin d'être neutres ou anodins, puisqu'ils auront une incidence non négligeable sur le déroulement de l'évaluation ainsi que sur les conclusions et les recommandations auxquelles parviendra l'évaluateur.

L'analyse des politiques nous enseigne que tous les acteurs ne peuvent pas s'organiser de la même manière en vue de défendre leurs intérêts et qu'ils ne disposent pas tous du même poids lorsqu'il s'agit de faire pression, voire d'influencer le processus décisionnel (Bachrach et Baratz, 1970). Il en va de même en ce qui concerne la définition d'un mandat d'évaluation. Les analystes de politiques considèrent qu'en observant les acteurs impliqués dans le processus de formulation d'un programme ou d'une politique, y compris le jeu de la distribution des ressources qui l'accompagne, il est possible d'en comprendre les résultats. Il s'agit là d'un élément intéressant à prendre en considération lorsqu'on compare les débats relatifs à la participation de la société civile aux processus décisionnels et d'élaboration de politiques en général (démocratie délibérative, panel de citoyens...) (Dryzek, 1990) avec les débats entourant le rôle de l'évaluateur et l'implication des parties prenantes, en particulier dans la littérature sur l'évaluation. Dans les deux cas, il y a une tendance qui plaide en faveur d'un renforcement de la participation et qui encourage un élargissement de l'inclusion des acteurs dans le processus décisionnel.

Les promoteurs de ce mouvement utilisent différents arguments pour appuyer leur point de vue. Selon eux, la participation au processus déci-

sionnel accroît la légitimité des choix effectués, et donc de l'intervention publique, et facilite la réussite de la mise en œuvre quant à la coopération, mais aussi au contenu, c'est-à-dire à la réalisation des objectifs définis précédemment. Il en va de même pour l'évaluation, où les promoteurs d'une démarche participative mettent l'accent sur le fait que la réalisation d'une évaluation peut être guidée par des considérations sociales, que la participation favorise l'utilisation des résultats produits et que l'inclusion d'un éventail plus large ou différencié de parties prenantes renforce la qualité, la validité et l'interprétation des données recueillies. Ce qui se reflète ultimement sur les conclusions et les recommandations formulées par l'évaluateur.

En résumé, même s'il existe des différences entre la définition d'un mandat évaluatif et l'élaboration d'une politique ou d'un programme du point de vue de la légitimité démocratique et de la reddition ultérieure de comptes, il est indéniable que leur contenu sera affecté par la nature des acteurs (impliqués, exclus ou qui ne souhaitent pas participer à cette étape, sur une base volontaire et parfois stratégique) et par les ressources dont ils disposent, et qui affectent la répartition des sources d'influence et de pouvoir. De plus, la définition d'un mandat évaluatif peut être fortement conditionnée par les pratiques existantes, les routines bureaucratiques et les formes d'institutionnalisation de la pratique, comme nous l'avons déjà vu. Il ne faut cependant pas conclure prématurément à un déterminisme dans la démarche évaluative. Ce serait négliger les dimensions partenariales de cette pratique. En effet, tout au long du processus, il y aura des moments de discussion, et parfois même de négociation, entre l'évaluateur et le commanditaire, en vue de concilier les dimensions scientifique et politique de l'exercice. L'évaluation étant davantage un métier (méthodologiquement exigeant) qu'une science, elle permet donc des aménagements. Ceux-ci ne sont pas exclusivement d'ordre politique. Ainsi, parallèlement à la diplomatie, l'évaluateur doit faire preuve de pragmatisme pour tenir compte des contraintes qui ne manquent jamais de surgir sur le plan des ressources (temporelles, budgétaires...). Comme nous allons le voir maintenant, la réalisation d'une évaluation est une alchimie entre le scientifiquement désirable et le matériellement réalisable.

La réalisation de l'évaluation : mise en œuvre

Les analystes de politiques considèrent que l'étape de la mise en œuvre consiste à agir en vue de concrétiser les objectifs d'une politique. Cette opérationnalisation des intentions en réalisations n'est pas un processus dénué d'enjeux. Il ne s'agit pas d'une simple transposition neutre et indifférente de la volonté des décideurs, puisque c'est au cours de cette étape que sont interprétées les directives gouvernementales (Muller, 2003). À ce sujet, les administrations se sont progressivement éloignées de la conception wébérienne selon laquelle la discrétion administrative est un obstacle à l'efficacité de la bureaucratie, à la neutralité et à l'anticipation de l'action administrative pour rencontrer les exigences de la nouvelle gestion publique. Pour les tenants de ce courant, la discrétion permet de maximiser l'efficacité des administrations en évitant la rigidité excessive qui caractérise bien souvent les organisations bureaucratiques.

Cette traduction des décisions politiques en effets concrets est bien souvent tributaire de ses créateurs et des autres acteurs concernés ultérieurement (Lemieux, 1995; Bardach, 1977). S'il peut y avoir plusieurs compréhensions et formulations d'un problème, il peut aussi y avoir des réinterprétations à l'étape de la mise en œuvre. Dans ce cas, les agents publics disposent d'une marge de manœuvre, voire d'une certaine latitude pour agir (Goggin *et al.*, 1990). Cette forme de discrétion administrative se retrouve en évaluation, où les acteurs impliqués dans l'exercice interprètent le mandat qui leur est confié en vue de s'adapter à des circonstances particulières (accessibilité des données, émergence d'une nouvelle question évaluative...).

Comme le démontre Bardach (1977), il est important de considérer le jeu entre les acteurs. Tous ne s'engagent pas (ou ne sont pas impliqués) de la même manière. L'enthousiasme à collaborer ou non à une évaluation est révélateur de la perception que les acteurs ont de cet exercice et, parfois, de leur anticipation des résultats. Certains tenteront donc d'orienter le processus à leur avantage, tandis que d'autres chercheront à l'entraver pour en compliquer la réalisation. À côté de cette attitude stratégiquement orientée, il est important de préciser que la littérature sur l'évaluation accorde une grande attention à l'implication de tiers. L'évaluation impliquant les parties prenantes (Cousins et Whitmore, 1998; Guba et Lincoln, 1989; Fetterman, 2001) permet de prendre en considération une variété

de points de vue et d'offrir à d'autres acteurs que le traditionnel duo commanditaire-évaluateur une possibilité d'influer sur la démarche évaluative. Comme nous l'avons déjà mentionné, l'identification préalable de toutes les parties prenantes en présence permet d'anticiper les situations auxquelles l'évaluateur risque d'avoir à faire face dans son travail.

En comparant la réalisation du mandat évaluatif à la mise en œuvre d'une politique ou d'un programme public, il n'est pas possible de faire abstraction des problèmes qui peuvent survenir au cours de cette étape et que la littérature sur l'analyse des politiques qualifie de « déficits de mise en œuvre » (Pressman et Wildavsky, 1973). Il s'agit des décalages, parfois considérables, qui se produisent entre les intentions des décideurs et les effets observés sur le terrain. Cela peut également se produire au cours d'une évaluation qui peut engendrer des conséquences inattendues. Dans les études d'analyse de politiques (Hogwood et Gunn, 1984; Sabatier, 1997), nous avons trouvé quelques éléments dont il faut tenir compte si l'on souhaite réduire ce risque et éviter que l'évaluation s'enlise, n'aboutisse pas ou produise des effets pervers. Il s'agit d'être attentif au fait qu'aucune contrainte n'est insurmontable, de favoriser l'implication d'une autorité reconnue, d'ancrer l'évaluation sur une théorie d'intervention valide, de définir des objectifs clairs, d'obtenir des informations pertinentes et de faire en sorte qu'une coordination parfaite soit établie entre les acteurs concernés.

Enfin, durant cette étape, il est nécessaire de veiller à la cohérence des actions entreprises. Par exemple, dans le cadre d'une évaluation d'envergure, il faut s'assurer que les différentes évaluations réalisées simultanément sur plusieurs sites (évaluation de multiples projets, de plusieurs sites...) sont cohérentes. Plus simplement, l'évaluateur peut, à chacune des étapes, s'interroger sur le sens des actions qu'il entreprend pour s'assurer qu'elles le mènent dans la direction souhaitée. Ces réflexions peuvent s'appuyer sur des standards de qualité, qui sont des guides utiles pour encadrer la démarche d'évaluation (par exemple, *The Joint Committee on Standard for Educational Evaluation*, Standards d'évaluation de la Société suisse d'évaluation – SEVAL). À ce sujet, nous considérons les standards comme des outils qui permettent à l'évaluateur de situer sa pratique à l'aune de normes ou de conceptions habituelles et usuelles en la matière, et non comme un catéchisme prescriptif à suivre à la lettre. En effet, bien souvent, l'évaluateur est amené à s'écarter de ces normes pour faire face

à certaines situations particulières ou tenir compte du contexte, éventuellement culturel, dans lequel il évolue. Dans de tels cas, il devra parfois adapter ses instruments de collecte ou ses grilles d'analyse, même si nous privilégions la parcimonie à ce sujet, car l'évaluateur peut avoir la sensation d'innover alors qu'il ne fait que réinventer la roue.

Du savoir produit à l'utilisation des conclusions : terminaison de la politique

Pour les évaluateurs, la production de connaissances sur le fonctionnement de l'État et sur les effets produits par son intervention est un objectif important. Ce constat est d'autant plus pertinent que les relations entre savoir et décision se renforcent et que l'expertise se popularise. Même si l'expertise publique a de tout temps joué un rôle important dans le processus décisionnel, les sociétés contemporaines font face à une offre d'expertise de plus en plus diversifiée provenant tant des universités que des groupes d'intérêt. Dans ce contexte, les décideurs publics sont amenés à justifier leurs choix. Cet impératif de transparence amplifie le recours à des données probantes (*evidence-based policy*). L'évaluation devient dès lors une source d'expertise non négligeable. C'est pour cette raison que les évaluateurs insistent sur l'impératif de rigueur méthodologique de leur démarche, afin de ne pas prêter le flanc à la critique, et donc de décrédibiliser le travail accompli.

La production de savoirs évaluatifs ne constitue cependant pas une fin en soi, et certains considèrent que le plus difficile commence lorsque le rapport d'évaluation est déposé. En effet, les connaissances produites peuvent éclairer les décisions ultérieures sur la politique ou le programme concerné. En prenant en considération les conclusions et les recommandations formulées dans l'évaluation, les décideurs ou les gestionnaires peuvent convenir du maintien, de la reformulation ou de la suppression de cette politique ou de ce programme. Il ne faut pas forcer le trait pour y voir un des enjeux majeurs du processus évaluatif. La question de l'utilisation des connaissances produites par les évaluations est centrale dans la littérature (voir le chapitre 8). Lors de chacune des étapes que nous avons mentionnées, l'évaluation peut susciter un processus d'apprentissage qui permet d'amender les buts et les moyens d'une politique ou d'un programme public. Certains considèrent d'ailleurs que le critère de vali-

dation d'une évaluation ne se résume pas uniquement aux questions de rigueur scientifique et méthodologique, mais qu'il dépend aussi fortement de l'utilisation des connaissances produites et de leurs répercussions sur le terrain. Il convient cependant de constater qu'il est rare, à l'exception de certains projets pilotes, qu'une évaluation entraîne la suppression d'une politique ou d'un programme.

Les chercheurs qui s'intéressent à l'évolution des politiques sur de plus ou moins longues périodes s'appuient sur plusieurs corpus théoriques qui peuvent être mobilisés par les évaluateurs en vue d'élargir le débat sur l'utilisation des conclusions et des recommandations d'une évaluation. Sur un axe partant de l'inutilisation des résultats jusqu'à leur utilisation, il est possible de mobiliser deux grandes approches : la dépendance au chemin emprunté (*path dependency*) et l'incrémentalisme. Selon la première approche, les ornières de la voie dans laquelle s'est engagée l'action publique sont si profondes qu'il est difficile de changer de direction (notamment quant aux coûts politiques ou financiers). Dans ce contexte, les évaluations seront peu utilisées. Dans une perspective axée davantage sur le changement, l'incrémentalisme nous enseigne que les aménagements de programmes publics sont les résultats d'un lent processus d'adaptation et de changements marginaux qui peuvent progressivement entraîner une reconfiguration importante des programmes. Dans ce cas de figure, les conclusions et recommandations d'une évaluation peuvent faciliter l'adoption de mesures de changements portant sur la mise en œuvre d'un programme (par exemple, moyens mis à disposition, processus), ce qui peut ensuite entraîner un réaménagement plus important de ce dernier.

On a pu voir que les débats et les enseignements théoriques de l'analyse des politiques peuvent enrichir la compréhension du contexte politique de l'évaluation. Cela implique qu'il ne faut pas réduire la pratique évaluative à un exercice « technocratique », mais plutôt qu'on peut la percevoir comme un élément du processus décisionnel. Toutefois, la littérature présentée dans ce chapitre nous rappelle que l'évaluation n'est pas la prise de décision. La grande variété des théories nous permettant de comprendre le cycle de la politique en général et les moments de la prise de décision en particulier nous rappelle sans cesse qu'une multitude de facteurs interviennent dans ces processus. Le savoir produit par l'évaluation n'est qu'une pierre (peut-être une clé de voûte) dans l'édifice du processus politique.

Pour aller plus loin :

ALKIN, Marvin C. (dir.) (2004), *Evaluation Roots. Tracing Theorists' Views and Influences*, Thousand Oaks, Sage Publications.

En donnant la parole aux théoriciens qui font évoluer la discipline de l'évaluation, Marvin Alkin offre une compilation de témoignages qui reflète la diversité des pratiques et dote le lecteur des principales clés de compréhension qui lui permettront de s'orienter et d'opérer des choix raisonnés au moment de lancer une démarche évaluative.

KUSEK, Jody Z. et Ray C. RIST (2006), *Dix étapes pour mettre en place un système de suivi et d'évaluation*, Montréal, Saint-Martin.

En s'appuyant sur une solide expérience de terrain acquise à la Banque mondiale, les auteurs présentent les différentes étapes du processus de suivi et d'évaluation. Les exemples proviennent de programmes de coopération au développement.

MASSARDIER, Gilles (2003), *Politiques et action publique*, Paris, Armand Colin.

Une synthèse de la littérature relative à l'analyse de l'action des pouvoirs publics qui s'adresse en priorité aux étudiants souhaitant se familiariser avec les concepts de base de l'analyse des politiques. En se concentrant sur les rôles des principaux acteurs impliqués dans la gestion publique, l'auteur contribue à la réflexion sur les mutations de l'État contemporain.

Internet

European Evaluation Society, <http://europeanevaluation.org>

International Development Evaluation Association, www.ideas-int.org

WWW Virtual Library: The World Wide Evaluation Gateway, www.policy-evaluation.org

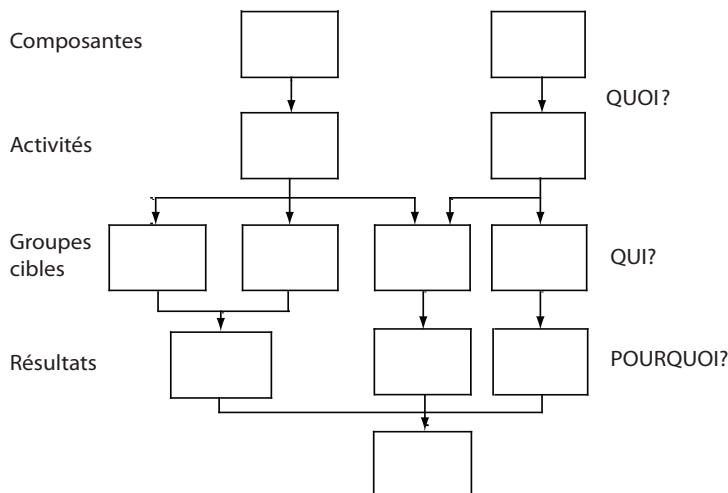
5

La construction du modèle logique d'un programme

Nancy L. Porteous

Le modèle d'analyse logique est un outil visuel visant à décrire un programme dans sa logique (ou sa théorie) fondamentale. Malgré leur diversité, les programmes ont en partage certains éléments, et le modèle logique est un diagramme de ces éléments communs qui illustre le contenu (quoi?), les destinataires (qui?) et la raison d'être (pourquoi?) du programme (voir la figure 5.1). Il est essentiel d'avoir une vision commune de la logique d'un programme avant d'élaborer des mesures de surveillance et d'évaluation.

Un modèle logique peut s'appliquer à des programmes de tout genre et dont la portée peut être large ou restreinte. On peut également élaborer ce type de modèle pour l'ensemble des activités d'un organisme. Dans le présent chapitre, le terme « programme » aura donc un sens assez large pour inclure les projets, les politiques ou d'autres types d'initiative tels que les interventions, les projets pilotes, les événements, les processus, les campagnes et les services. Un programme peut être très vaste ou très restreint, ou encore se situer quelque part entre ces deux extrêmes. Fondamentalement, un programme consiste en une série d'activités appuyées par un ensemble de ressources et visant à atteindre des résultats spécifiques au sein de groupes cibles définis.

FIGURE 5.1**Modèle d'analyse logique**

Il est utile de définir ici les termes qu'on utilise dans un modèle logique pour décrire les principaux éléments d'un programme :

Quoi?

- Les **composantes** d'un programme sont des séries d'activités étroitement apparentées.
- Les **activités** sont les mesures mises en œuvre par le programme pour parvenir aux résultats désirés.

Qui?

- Les **groupes cibles** sont les clientèles – individus, groupes, collectivités ou organismes à qui s'adresse le programme.

Pourquoi?

- Les **résultats** sont les changements que l'on espère effectuer par l'entremise du programme.

Éléments facultatifs :

- les **intrants** sont les ressources qui permettent la réalisation des activités;
- les **extrants** sont les services, les événements ou les biens produits par les activités;
- les **facteurs contributifs** représentent ce qui explique que les activités ou les extrants produisent les résultats (par exemple, les réactions des participants, comme leur intérêt ou leur satisfaction à l'égard du programme);
- les **facteurs externes** sont des éléments situés à l'extérieur de la sphère du programme, mais qui exercent une influence sur ce dernier.

Nous traiterons de tous ces éléments, mais abordons tout d'abord les notions de mise en situation et de contexte.

Utilité, avantages et défis des modèles logiques

Le modèle logique de programme constitue un outil important de planification et d'évaluation depuis la fin des années 1980. Ce modèle permet de compiler et de visualiser l'information concernant un programme à l'aide d'un schéma, afin de favoriser la compréhension des enjeux du programme. Idéalement, l'élaboration d'un modèle logique s'effectue à l'étape de la planification d'un programme. S'il n'existe aucun modèle logique, l'élaboration d'un tel outil devient une première étape cruciale dans tout processus d'évaluation.

Les modèles logiques sont également très utiles pour communiquer les tenants et aboutissants d'un programme à divers intervenants à l'égard d'une foule d'objectifs. Un modèle logique permet de transmettre rapidement la raison d'être d'un programme au moment de l'initiation et de la formation du personnel et des bénévoles (voir le tableau 5.1).

Le processus d'élaboration d'un modèle logique a en soi une valeur inestimable. Les dialogues, les discussions et parfois même les débats qui s'ensuivent permettent d'élucider des conceptualisations nouvelles ou encore d'autres interprétations possibles d'un programme et des objectifs qu'il est censé atteindre. Nous l'avons déjà mentionné, un modèle logique dûment établi constitue un outil très utile de planification et d'évaluation. Le tableau 5.2 en énonce les avantages.

TABLEAU 5.1**Possibilités d'utilisation du modèle logique d'un programme**

Buts	Publics visés
Planification	Gestionnaires, employés, partenaires et autres intervenants, planificateurs
Demandes de subventions	Bailleurs de fonds
Communication	Décideurs, cadres supérieurs, collègues d'autres organismes, partenaires du programme, représentants des médias, participants
Initiation et formation	Nouveaux employés ou bénévoles
Surveillance et évaluation	Spécialistes de l'évaluation et intervenants

TABLEAU 5.2**Avantages liés à l'élaboration et à l'utilisation d'un modèle logique de programme**

Processus : élaborer un modèle logique	Produit : disposer d'un modèle logique complet
Comblar le fossé entre la planification stratégique et la planification opérationnelle.	Résumer les principaux éléments d'un programme (de préférence sur une seule page).
Permettre aux intervenants de discuter du programme et de s'entendre sur la description qu'ils en font.	Expliquer la théorie qui sous-tend les activités du programme de même que les hypothèses.
Ouvrir la voie à des façons différentes ou novatrices d'élaborer un programme.	Pouvoir décrire facilement le programme à d'autres personnes.
Faire ressortir les points de compréhension ou les perceptions qui diffèrent à l'égard du programme.	Démontrer les relations causales entre les activités et les résultats (par exemple, telle activité est censée produire tels résultats).
Expliquer nettement la différence entre les activités et les résultats escomptés.	Aider à déterminer qui est responsable de tels résultats au cours de telle période.
Faciliter la détermination de questions importantes pour l'évaluation.	Contribuer à l'élaboration de mesures de rendement du programme pouvant être évaluées de façon suivie.

Malgré ces points forts, le modèle logique comporte aussi certains inconvénients :

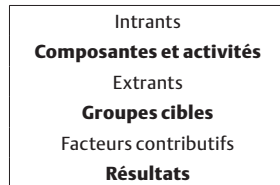
- L'utilisation d'un modèle logique peut exiger du temps.
- Il ne s'agit que d'une représentation de la réalité – un simple modèle :
 - On ne saurait étudier tous les volets d'un programme.
 - On risque de trop simplifier le programme, par exemple en omettant d'illustrer de façon adéquate le contexte du programme et les influences externes qui agissent sur lui.

- On risque de transformer des systèmes complexes en systèmes linéaires.

À propos des modèles logiques, Westley, Zimmerman et Patton (2006) énoncent cet avertissement général : « Les conceptualisations de ce genre présupposent une connaissance préalable du déroulement exact d'une innovation. Dans les milieux complexes et très dynamiques, les préacquis ne sont pas possibles ni souhaitables, puisqu'ils peuvent restreindre l'ouverture et la souplesse » (p. 237; notre traduction).

Les éléments d'un modèle logique

Les éléments essentiels d'un modèle logique sont les composantes et les activités qui s'y rattachent, les publics cibles et les résultats à court, moyen et long terme. Dans certains cas, il peut également être utile d'inclure les éléments facultatifs que sont les intrants, les extrants et les facteurs contributifs.



Composantes, activités et groupes cibles

Les **composantes** d'un programme sont des ensembles d'activités étroitement liées. Leur nombre dépend de l'envergure du programme et de la façon dont les intervenants le conceptualisent et l'administrent. Le modèle logique applicable à un vaste programme pourrait comporter plusieurs composantes, mais un petit programme peut n'en comporter qu'une seule.

On les exprime le plus souvent à l'aide d'énoncés courts, comme dans les exemples suivants :

- défense des droits
- développement communautaire
- collecte de fonds
- éducation publique
- consultation
- intervention immédiate
- service de référence
- formation

Conseil : S'attarder aux thèmes ou aux ensembles d'activités. Il s'agira parfois des grandes stratégies utilisées ou encore des secteurs d'activités ou de services.

Les **activités** sont les principales fonctions ou tâches que le personnel du programme accomplit, ou encore les services qu'il offre. Elles représentent le moyen d'atteindre les résultats escomptés. Pour aider les intervenants à y réfléchir, réunissez tous les documents qui portent sur le programme. Il peut s'agir d'une description des activités du programme, de plans de travail ou de plans opérationnels.

On les désigne par des verbes d'action, comme :

- animer
- élaborer
- fournir
- établir
- distribuer
- former
- identifier
- soutenir

Conseil: Éviter les aspects administratifs du programme comme le service de la paie ou les évaluations de rendement.

Les **groupes cibles** sont les individus, les groupes, les organismes et les collectivités auxquels sont destinés les services prévus par le programme, c'est-à-dire les clients, les destinataires, les bénéficiaires, les consommateurs, les populations jugées prioritaires ou le public « cible » du programme. Ils peuvent être définis en fonction de caractéristiques sociodémographiques (comme l'âge, le revenu, l'emploi, la situation géographique, la composition du ménage, l'éducation, le sexe, la langue, un handicap, l'origine ethnique ou l'orientation sexuelle) ou en fonction de la santé ou du statut social, de problématiques particulières ou de troubles de comportement.

Les groupes cibles se définissent à l'aide de descriptions brèves :

- femmes autochtones
- familles à faible revenu vivant en milieu rural
- nouveaux immigrants, réfugiés
- aînés vivant seuls
- sans-abri ou personnes à risque de le devenir
- jeunes d'origine haïtienne de 14 à 17 ans

Conseil: Faire preuve d'une grande précision en combinant plusieurs caractéristiques.

Les résultats

Les résultats sont les changements que le programme devrait produire au sein de chaque groupe cible, autrement dit ils renvoient aux objectifs, aux buts, aux solutions, aux avantages, aux effets ou aux conséquences anticipés de la mise en œuvre du programme. Il s'agit de choisir la terminologie qui convient le mieux au contexte de chaque programme. Il reste que les résultats portent sur les effets du programme plutôt que sur les activités qu'il comporte ; il s'agit donc de la finalité du programme et non de la méthode ou du processus employé. Les résultats répondent à la question « Et alors ? ».

Le nombre de résultats varie selon l'envergure et la complexité du programme, et en fonction de ses objectifs et de son public cible. Par la mise en œuvre d'un programme, on espère généralement des résultats positifs. Par contre, alors que certains programmes viseront à entreprendre ou à maintenir un changement positif, d'autres contribueront plutôt à ce que la situation des participants se détériore moins rapidement.

Il est important de bien démontrer la séquence des changements souhaités par le groupe cible. En général, on obtient non pas une solution unique résultant de l'addition des activités et des intrants d'un programme, mais bien une série de résultats, l'un menant au suivant. On peut exprimer cette séquence par des phrases comme : « **Si** tel résultat se produit, **alors** tel autre résultat devrait en découler, entraînant tel autre résultat. » On peut parler d'une hiérarchie ou d'une séquence de résultats.

TABLEAU 5.3

Séquence typique de résultats

Résultats à court terme	Changements relatifs au degré de sensibilisation à une problématique, aux connaissances, aux aptitudes ou aux compétences
↓	
Résultats à moyen terme	Changements d'habitudes ou de comportements
↓	
Résultats à long terme	Changements des conditions sociales, économiques, environnementales ou liées à la santé

Les résultats à court terme sont les effets directs d'un programme sur les participants, par exemple, les changements touchant le degré de sensibilisation à l'égard d'une question ou d'un problème, l'accroissement des connaissances ou un changement d'attitude ou le perfectionnement de

compétences. Les résultats à moyen terme se traduisent par un changement dans les habitudes, dans les comportements ou dans les aptitudes. Les résultats à long terme sont les conséquences sociales, économiques ou environnementales du programme, de même que ses répercussions sur l'ensemble de la collectivité. Mais quel que soit le programme, il y a généralement fort à parier que ces résultats à long terme seront peu nombreux.

Ces trois ordres de résultats s'articulent donc de façon séquentielle. Par exemple, **si** un programme d'habiletés parentales offre des sessions de groupes conçues à l'intention de jeunes parents d'enfants d'âge préscolaire, **alors** ces parents pourront accroître leurs connaissances et leurs aptitudes pour s'occuper de leurs enfants. **Si** les parents approfondissent leurs connaissances et leurs aptitudes afin de mieux s'occuper de leurs enfants, **alors** ils auront des comportements plus appropriés en tant que parents, comme préparer des repas équilibrés, prodiguer de bons soins aux enfants malades, communiquer de façon efficace, fixer des limites,

FIGURE 5.2
Modèle logique d'un programme d'habiletés parentales

Composantes	Éducation en matière de santé
↓	
Activités	• Animer des discussions à partir de modules rédigés sur les sujets relatifs aux habiletés parentales • Distribuer des dépliants sur divers sujets • Distribuer des dépliants sur d'autres ressources communautaires
↓	
Groupes cibles	Parents d'enfants âgés entre 2 et 4 ans, en particulier
↓	
Résultats à court terme	Les parents qui participent auront : • Une connaissance accrue des soins à prodiguer à un jeune enfant • Un soutien accru et constant des pairs • Une connaissance accrue des ressources et des services offerts • Des habiletés parentales améliorées
↓	
Résultat à moyen terme	Nombre accru de parents qui adoptent des comportements parentaux appropriés
↓	
Résultat à long terme	Nombre accru d'enfants capables d'atteindre leur niveau optimal de développement physique, mental, émotionnel et social

aborder les questions de sexualité, ou favoriser l'estime de soi chez leurs enfants. **Si** les parents sont capables d'aborder leur rôle de façon plus saine, **alors** leurs enfants devraient être en mesure d'atteindre un niveau optimal en termes de développement physique, mental, émotionnel et social. En d'autres mots, leurs enfants seront plus sains.

Il est à noter que les activités d'un tel programme sont destinées aux parents. Ce sont eux qui forment le groupe cible, celui des destinataires des services. Toutefois, leurs enfants en sont, en bout de ligne, les bénéficiaires. Voilà pourquoi il est important de préciser dans l'énoncé des résultats qui bénéficie des changements induits par chacun de ces résultats.

La terminologie utilisée pour décrire les résultats à court, moyen et long terme varie selon les organismes et les secteurs. Répétons-le, il est important de choisir la terminologie qui s'accorde le mieux au contexte de l'étude.

TABLEAU 5.4

Exemples de terminologies pour des séquences de résultats

	Bibliographie	Gouvernement du Canada
Résultats à court terme	Résultats	Résultats immédiats (ou directs)
Résultats à moyen terme		Résultats intermédiaires (ou indirects)
Résultats à long terme	Impacts	Résultats finaux (ou ultimes)

La distinction entre les résultats à court, moyen et long terme relève davantage de la séquence que de la durée. L'expression court terme n'équivaut pas nécessairement à un mois, et parler de long terme ne veut pas toujours dire cinq ans. Tout dépend du programme et de ses objectifs. Par exemple, il aura fallu près de 25 ans pour constater les résultats d'un programme multistratégique visant à contrer le tabagisme. À l'inverse, les effets à long terme d'un programme d'immunisation visant à réduire le taux de maladies contagieuses se manifestent beaucoup plus rapidement.

Les éléments facultatifs d'un modèle logique

Les **intrants** sont les ressources qui permettent aux activités d'un programme d'être mises en œuvre. On peut compter parmi les intrants les ressources financières (l'encaisse, le budget, les fonds), les ressources humaines (le personnel, les bénévoles, les experts techniques), les ressources physiques (les installations) ou les ressources matérielles (l'équipement et les matériaux).

Les **extrants** sont les services, les événements ou les biens produits par les activités des programmes à l'intention des groupes cibles. Ils sont tangibles et quantifiables.

Les **facteurs contributifs** représentent ce qui mène les activités ou les extrants vers les résultats à court terme. La participation d'un groupe cible à un programme est un bon exemple de facteur contributif typique. La réaction des participants à l'égard du programme en constitue un autre exemple. Les participants sont-ils intéressés, enthousiastes? Sont-ils satisfaits? Une activité, ou un extrant, ne pourra vraisemblablement pas produire les résultats à court terme désirés si les participants n'en sont pas satisfaits ou s'ils n'y trouvent aucun intérêt. Cela dit, l'intérêt et la satisfaction ne sont pas les buts recherchés; ce sont plutôt les moyens par lesquels les résultats peuvent être atteints.

Les **facteurs externes** sont des éléments situés à l'extérieur de la sphère du programme, mais qui exercent une influence sur lui. À titre d'exemple, pensons à des changements dans la situation politique ou économique, à l'agitation civile, aux médias, à des événements imprévus tels que des épidémies ou des désastres naturels, etc. Ces facteurs externes illustrent la complexité potentielle des contextes des programmes.

Influence et responsabilité

Plus on se dirige vers le bas dans le schéma d'un modèle logique, moins l'influence qu'exerce un programme sur l'atteinte réelle de ses résultats est grande. En d'autres termes, lorsqu'on passe des intrants aux extrants, il est de plus en plus probable que des forces externes au programme interviennent. Steve Montague, du Réseau de gestion de la performance, distingue quatre degrés de contrôle ou d'influence.

1. Contrôle direct. Les programmes sont responsables de la gestion de leur budget et des activités visant à produire les extrants. Même s'il y a des changements dans l'environnement d'un programme, y compris la réaffectation des ressources, les gestionnaires et le personnel d'un programme doivent « contrôler » les aspects opérationnels de ce programme, notamment les activités et les extrants.
2. Influence directe. Les programmes *influencent* l'atteinte réelle des résultats, mais ils ne peuvent pas contrôler certains éléments, comme le fait de rejoindre ou non les groupes cibles, leurs réactions aux activités, ou encore le fait de constater des changements immédiats par suite de leur participation aux programmes. L'influence des programmes, dans ces cas, est dite *directe* – les équipes des programmes adaptent intentionnellement leurs activités en vue de satisfaire aux besoins particuliers des groupes cibles.
3. Influence indirecte ou concourante. Lorsqu'il s'agit de résultats à moyen et à long terme, l'influence d'un programme est, au mieux, indirecte. Les programmes, ainsi que tous les autres facteurs touchant les participants au programme, peuvent *contribuer* à des changements dans leurs habitudes ou leur comportement.
4. Hors du contrôle ou de l'influence. De nombreux facteurs externes interviennent dans la vie des participants à un programme et influent sur la manière dont ils prennent part au programme et y réagissent. Bien que les programmes prévoient et reconnaissent ces facteurs, ils ne peuvent en général pas les contrôler ni les influencer.

Le tableau 5.5 illustre le degré d'influence normalement associé à chaque élément d'un programme. À mesure que le degré de contrôle ou d'influence diminue, le niveau de responsabilité que l'on attend normalement de la part des gestionnaires des programmes et de leurs équipes décroît.

En ajoutant des cercles concentriques autour du modèle logique, on arrive à illustrer les degrés d'influence décroissants sur la séquence des résultats. À titre d'exemple, voyons le cas du Centre de prévention et de contrôle des maladies chroniques, qui provient de l'Agence de la santé publique du Canada.

TABLEAU 5.5

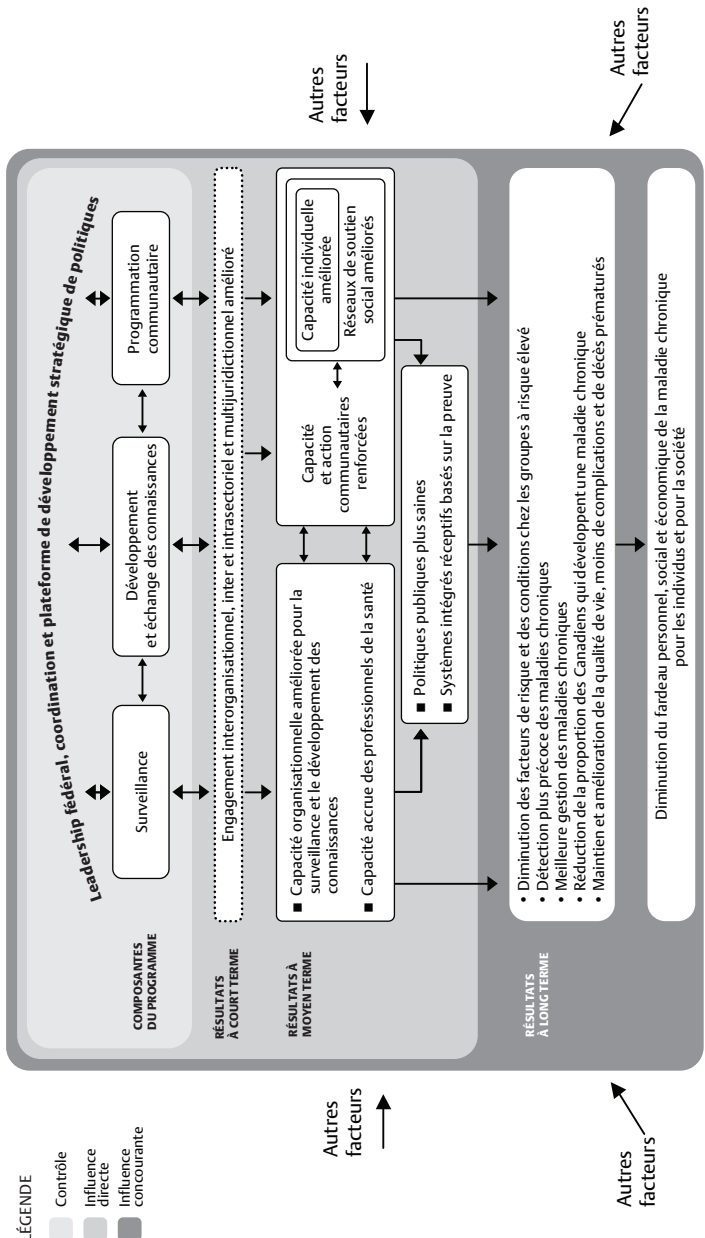
**Degrés d'influence des éléments d'un programme
et niveaux de responsabilité attendue**

Degré de contrôle ou d'influence	Éléments du programme	Niveau de responsabilité imputé au programme
Contrôle direct ↓	Intrants : ressources qui supportent les activités	Élevé
	Composantes et activités qui s'y rattachent : réalisations, services rendus	
	Extrants : biens, services ou événements	
Influence directe ↓	Groupes cibles : individus, groupes ou collectivités	Modéré
	Facteurs contributifs : participation, réactions	
	Résultats à court terme : changements relatifs au niveau de sensibilisation, aux connaissances, aux aptitudes ou aux compétences	
Influence concourante ou indirecte ↓	Résultats à moyen terme : changements relatifs aux habitudes et au comportement	Faible
	Résultats à long terme : changements dans les conditions sociales, économiques, environnementales ou sanitaires	
Hors du contrôle ou de l'influence	Facteurs externes : changements dans la situation politique ou économique, agitation civile, médias, événements imprévus tels que des épidémies ou des désastres naturels	Aucun

La figure 5.3 représente un modèle logique de niveau élevé qui ne renvoie qu'aux résultats attribuables aux quatre composantes principales de l'organisme. Par la suite, on peut établir un modèle logique plus détaillé incluant des sous-composantes, des activités, des extrants et des résultats à court et moyen terme plus précis et correspondant à chacune des composantes. Il s'agit d'une approche hiérarchique dont le but est de créer des modèles logiques comportant des niveaux de détail différents selon les destinataires et les objectifs des programmes.

FIGURE 5.3

Modèle logique du Centre de prévention et de contrôle des maladies chroniques



L'élaboration d'un modèle logique

Qui devrait participer ?

Traditionnellement, les évaluateurs procédaient pratiquement seuls à l'élaboration d'un modèle logique en guise de point de départ du processus d'évaluation. Ils se fiaient à la documentation existante relative au programme et aux quelques entretiens qu'ils obtenaient avec les principaux membres du personnel. De nos jours, il est beaucoup plus commun, voire recommandé, de favoriser la participation des intervenants du programme dans l'élaboration des modèles logiques. Bien souvent, le processus d'élaboration d'un modèle logique est aussi utile que le modèle logique lui-même; et cela s'applique autant aux intervenants du programme qu'aux évaluateurs.

Il existe plusieurs façons de faire participer les intervenants à l'élaboration du modèle logique. La méthode à privilégier dépend des relations entre les intervenants, de la complexité du programme, du délai disponible et des habiletés de l'évaluateur en tant qu'animateur. Trois grandes options sont offertes :

1. Préparer une ébauche pour amorcer la discussion avec les intervenants (on peut ainsi gagner du temps, mais les intervenants pourraient se sentir moins concernés).
2. Partir de zéro avec l'ensemble du groupe (les participants s'impliquent plus, mais cela exige plus de temps).
3. Demander à des sous-groupes de préparer des ébauches sur différents éléments du programme (on peut ainsi atteindre l'équilibre entre l'implication du groupe et le temps d'exécution nécessaire).

Quelle que soit la méthode utilisée, il faut demander aux principaux intervenants de vérifier l'exactitude des travaux; reconnaître la nécessité de contester ou de remettre en question la logique du programme; et obtenir un accord sur le modèle logique établi avant de procéder à l'évaluation.

L'élaboration d'un modèle logique est un processus itératif. Il faudra sans doute passer par plusieurs ébauches avant de parvenir à la version finale. Il suffit de mettre l'accent sur les avantages du processus et de ne pas se soucier de produire le modèle logique parfait dès le départ.

L'évaluateur devrait stimuler la discussion parmi les intervenants s'il est confronté à des perspectives hautement divergentes quant à la façon dont le programme devrait être traduit dans le cadre du modèle logique. Il devrait s'efforcer de comprendre les hypothèses, les valeurs et les partis pris des intervenants et tenter de trouver un terrain d'entente. S'il est impossible d'obtenir un consensus, il faut déterminer si tous les intervenants peuvent au moins accepter une version particulière du modèle logique. En dernier recours, il est possible d'utiliser plus d'un modèle logique pour tenir compte de tous les principaux éléments du programme au moment de la planification et de la détermination des questions et des indicateurs de l'évaluation.

Il faut parfois plusieurs essais avant d'en arriver à un modèle réduit à sa plus simple expression. Il peut être utile de modifier l'ordre des composantes ou des groupes cibles pour atténuer « l'effet spaghetti », soit un enchevêtrement qui réduit considérablement l'impact visuel du modèle et qui ne permet pas de bien saisir la logique du programme.

Par où commencer ?

Il n'existe pas de bon ou de mauvais point de départ. Tout dépend de la participation de chacun au programme et de la place qu'il y occupe. Les exécutants de première ligne jugent souvent plus facile de commencer par les activités et les groupes cibles, tandis que les gestionnaires préfèrent habituellement commencer par les résultats escomptés. La décision peut cependant dépendre de l'étape de développement du programme. Dans le cas de nouveaux programmes, il est souvent plus approprié de commencer par les résultats visés (quels changements espère-t-on obtenir ?) pour ensuite examiner les moyens, alors qu'il est plus naturel s'il s'agit de programmes déjà bien établis de définir d'abord les activités (que fait le programme ?) avant d'évaluer les effets.

Dans le cas de programmes établis, il faut consulter toute la documentation disponible, y compris :

- les plans stratégiques de l'organisme (sa vision, sa mission, son mandat et ses valeurs) ;
- les organigrammes ;
- les plans opérationnels, budgétaires et d'exécution relatifs au programme ;

- le matériel d'initiation et de formation du personnel;
- le matériel publicitaire produit pour le programme (les brochures et circulaires);
- les rapports d'évaluation des comités antérieurs, etc.

Il est toujours préférable de se concentrer sur la façon dont le programme fonctionne *de facto*, et non sur la façon dont il a été conçu, sur son fonctionnement antérieur ou sur son déroulement idéal.

Dans le cas de nouveaux programmes, on se réfère généralement :

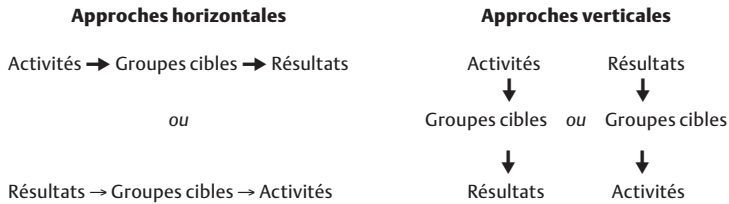
- aux évaluations des besoins et aux enquêtes préalables effectuées pour le programme;
- aux descriptions ou aux rapports d'évaluation de programmes similaires offerts par d'autres organismes;
- aux rapports de consultation des intervenants sur le sujet à l'étude.

Le format et le sens du déroulement

Il existe une foule de guides et de manuels sur les modèles logiques, et autant de termes et de formats. Il faut éviter de s'y attarder pour se concentrer plutôt sur la compréhension de la conception des modèles logiques et choisir l'approche qui convient le mieux au contexte du programme à évaluer. Il est important d'utiliser le vocabulaire et les méthodes employés par chaque organisme ou par plusieurs organismes qui partagent régulièrement de l'information sur leurs programmes.

Certains modèles sont présentés sous forme de tableaux ou de matrices, mais les tableaux traduisent difficilement les relations de cause à effet présumées entre les activités et les résultats. Par conséquent, on recommande fortement d'encourager les équipes associées aux programmes à *tracer un schéma* du modèle logique de leur programme. Une représentation visuelle a toujours un impact important.

Le mode traditionnel d'élaboration de modèles logiques consiste en l'usage de boîtes reliées par des flèches horizontales ou verticales (voir la figure 5.1). Dans un modèle logique vertical, on retrouve donc les activités en haut de la page et les résultats en bas de page, ou inversement. Dans le cas d'un modèle horizontal, les activités peuvent se situer à droite et les résultats à gauche, ou vice-versa :

FIGURE 5.4**Deux formats de modèles logiques**

On peut utiliser divers progiciels pour tracer le modèle logique définitif. Les équipes de programmes pourraient plus facilement utiliser des logiciels spécialisés pour la création de tableaux et de diagrammes ou d'élaboration de processus, mais elles n'en auraient peut-être pas un usage très fréquent. Les applications de bureautique servant à produire des présentations seront sans doute plus facilement accessibles. Certains utilisent des logiciels de traitement de texte ou des chiffriers électroniques, mais ces derniers peuvent être d'usage complexe.

Dans certaines situations, il se peut que le processus d'élaboration du modèle logique et son format doivent être revus et adaptés. Michael Quinn Patton rappelle ainsi que :

Parmi les groupes culturels où la tradition orale est très présente, comme c'est le cas chez des peuples autochtones d'Amérique du Nord [...], le langage et la méthode utilisés par les évaluateurs pour établir des modèles logiques [...] peuvent être très rebutants. Les modèles logiques sont typiquement linéaires et rigides, et le vocabulaire théorique est parfois intimidant. En lieu et place, il est possible d'adapter le modèle logique en fonction des modes de communication culturellement préférables, cela afin de mettre en valeur les connaissances et savoir-faire des populations [...]. [On peut ainsi] utiliser le scénario-maquette, qui permet d'organiser les récits en une séquence d'images plutôt que de recourir aux flèches et aux cases qu'affectionnent les concepteurs d'organigrammes. (Patton, 2002; notre traduction)

L'évaluation des modèles logiques

La représentation d'un programme n'est ni bonne ni mauvaise en soi, mais certaines représentations sont nettement plus fiables sur le plan théorique. Il s'agit d'amener les gestionnaires et le personnel des programmes à s'interroger sur la logique qui sous-tend leurs programmes et

à en justifier les relations sous-jacentes de cause à effet à l'aide d'éléments probants puisés dans la littérature.

Voici une liste de questions pouvant servir à l'évaluation de modèles logiques. Si la réponse à l'une de ces questions est négative, les intervenants devraient prendre le temps de repenser le programme et son modèle logique.

TABLEAU 5.6

Liste de vérification pour l'évaluation d'un modèle logique

Pertinence et réalisme
☐ Les activités et les résultats répondent-ils à un besoin manifeste au sein du groupe cible ? ☐ Les ressources sont-elles suffisantes pour entreprendre les activités du programme ? ☐ Les facteurs externes, y compris les imprévus, ont-ils été pris en compte ?
Précisions suffisantes
☐ Tous les éléments sont-ils clairement énoncés ? ☐ Toutes les activités, tous les groupes cibles et tous les résultats prévisibles sont-ils inclus ? ☐ Existe-t-il un document écrit pour accompagner l'élément visuel ?
Exactitude
☐ Y a-t-il des erreurs ou des omissions ? ☐ Les résultats inscrits sont-ils de véritables résultats, et non pas des extrants ou des activités ? ☐ Tous les résultats annoncent-ils un changement projeté ? ☐ Les perspectives des principaux intervenants ont-elles été prises en compte ? ☐ Les intervenants sont-ils à l'aise avec le modèle logique ?
Logique
☐ Existe-t-il des lacunes, c'est-à-dire toutes les composantes conduisent-elles à un ou à plusieurs résultats par l'entremise d'activités et de groupes cibles (bref, manque-t-il quelque chose) ? ☐ La séquence des résultats est-elle logique ? Est-il raisonnable de prévoir que les activités du programme aboutiront réellement aux résultats escomptés ? En d'autres mots, les relations causales sont-elles plausibles et corroborées par la documentation, l'expérience de l'équipe du programme ou celle d'autres personnes assurant la prestation de programmes semblables ? Si vous dirigez <i>telles</i> activités vers <i>tels</i> groupes cibles, est-il alors raisonnable de prévoir obtenir <i>tels</i> résultats ? ☐ Les résultats du programme sont-ils en harmonie avec la vision, la mission, les principes et le plan stratégique à long terme de l'organisme ? ☐ Les résultats seront-ils mesurables ? ☐ L'équipe du programme est-elle à l'aise pour accepter la responsabilité des résultats à court terme ?
Simplicité visuelle
☐ Le modèle logique tient-il en une seule page ? ☐ Le modèle logique se lit-il facilement en un coup d'œil ? ☐ A-t-on évité la multiplication abusive des cases ou un enchevêtrement de type « spaghetti » ?
Source : Porteous <i>et al.</i>, 2002. Tiré en partie de Poole <i>et al.</i> (2000), « Evaluating performance measurement systems in nonprofit agencies: The Program Accountability Quality Scale [PAQS] », <i>American Journal of Evaluation</i>, vol. 21, n° 1, p. 15-26.

Le modèle logique est un outil très utile de planification et d'évaluation. Il est en effet important de bien cerner la théorie d'un programme avant de commencer à mesurer son efficacité et son rendement. La flexibilité du modèle logique, tant en termes de terminologie que dans la grande diversité des formats qu'il est possible d'employer, permet de l'adapter à n'importe quel programme, quelle qu'en soit l'envergure.

Pour aller plus loin :

PORTEOUS, N. L., B. J. SHELDRIK et P. J. STEWART (1997), *Trousse d'évaluation des programmes*, Ottawa, Service de santé d'Ottawa-Carleton.

Un guide indiquant, étape par étape, comment élaborer un modèle logique à l'aide de feuilles de travail et d'exemples, de même qu'un article portant sur l'enseignement de cette technique. Commandes en ligne : www.ottawa.ca/residents/funding/toolkit/index_fr.html.

PORTEOUS, N. L., B. J. SHELDRIK et P. J. STEWART (2002), « Présentation des modèles logiques aux équipes de projet pour faciliter le processus d'apprentissage », *Revue canadienne d'évaluation de programmes*, Édition spéciale, vol. 17, n° 3, p. 113-141.

Internet

Le Centre for Disease Control des États-Unis a produit un bon guide d'élaboration et d'utilisation des modèles logiques : www.cdc.gov/DHDSP/state_program/evaluation_guides/logic_model.htm

La Fondation W. K. Kellogg a publié en 2002 un document très complet intitulé *Using Logic Models to Bring Together Planning, Evaluation and Action: Logic Model Development Guide* : www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf

Promotion Santé Suisse a élaboré un modèle de classification de résultats qui constitue une ressource utile pour la création de modèles logiques de programmes de santé publique : www.gesundheitsfoerderung.ch/common/files/knowhow/tools/N107713_Ergebnismodell_fr.pdf

Le Secrétariat du Conseil du Trésor du Canada offre un outil en ligne sur la gestion axée sur les résultats (GAR). Le second module porte sur la description d'un programme, d'une politique ou d'un projet et sur l'élaboration du modèle logique s'y rattachant : www.tbs-sct.gc.ca/eval/tools_outils/RBM_GAR_cour/Enh/module_o2/module_o2_f.asp

L'Unité de développement et d'évaluation des programmes du Programme d'extension des coopératives de l'Université du Wisconsin a un excellent site Internet sur les modèles logiques, où l'on trouve un module d'auto-information, des feuilles de travail, une bibliographie et plusieurs exemples : www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html

6

Attribution et causalité des effets

Robert R. Haccoun et Pierre McDuff

Soucieuse du fait que la transmission de maladies graves comme le VIH et l'hépatite est favorisée par l'échange de seringues contaminées, une importante municipalité canadienne décide d'autoriser et de financer un programme d'échange de seringues destiné aux héroïnomanes. Une évaluation effectuée à la suite de la mise en œuvre de ce programme démontre que l'incidence de nouveaux cas d'infection (VIH et hépatite) est statistiquement inférieure à ce qu'elle était auparavant. Quelque chose a donc changé! Mais pouvons-nous conclure hors de tout doute que ce programme d'échange est la cause de cette diminution et qu'il est donc un « succès »? D'autres villes devraient-elles suivre cet exemple?

Les réponses à ces questions et l'assurance avec laquelle nous pouvons les exprimer dépendent principalement des démarches et des procédures qui ont été mises en œuvre pour évaluer ce programme. Parmi ces démarches, le choix du devis d'évaluation, la constitution du groupe témoin et les menaces aux validités interne et externe constituent des éléments importants de l'attribution causale et seront abordés plus en détail dans ce chapitre.

Bien qu'il existe une vaste gamme de techniques évaluatives, celles-ci partagent toutes – en particulier celles qui impliquent des personnes – un certain nombre de caractéristiques :

- la mesure d'au moins un phénomène qui devrait logiquement être influencé (la variable dépendante) par le traitement ou l'intervention (la variable indépendante);
- l'étude d'un ou de plusieurs groupes de personnes ciblées par l'intervention, le traitement ou le programme et (ou) qui y participent;
- (au moins) une collecte de ces informations. Ces collectes sont organisées d'une manière précise que nous nommons « devis d'évaluation ». Ces devis peuvent être très simples ou très complexes, et le choix de l'un d'eux dépend du degré de précision requis et du niveau d'inférence requis, aussi bien que des possibilités de réalisation concrète de l'évaluation. Certains devis ne permettent de faire qu'une simple description des résultats, alors que d'autres peuvent fournir des informations sur la ou les causes ayant mené à ces résultats. Dans ce chapitre, nous nous concentrerons sur les évaluations causales;
- enfin, toute évaluation exige une analyse, quantitative ou qualitative, des informations.

La différence entre l'évaluation et l'évaluation causale

Si l'*évaluation de programmes* permet de quantifier un résultat (par exemple, le taux d'infections a-t-il véritablement diminué?) et de le qualifier (cette différence est-elle grande ou petite?), l'*évaluation causale* sert quant à elle à estimer la cause probable de ce résultat (est-ce bien le programme d'échange de seringues qui est la cause de ce résultat?).

En évaluation causale, la cause est nommée *variable indépendante*, alors que l'effet prend le nom de *variable dépendante*. Ainsi, lorsqu'une boule de billard blanche percute une boule noire, la trajectoire de la boule noire (la variable dépendante) est l'*effet* que la boule blanche (la variable indépendante) *cause*. En évaluation causale, la ou les *variables indépendantes* sont les interventions qui définissent le programme. Ainsi, dans l'exemple du programme d'échange de seringues, la variable indépendante est la présence ou non du programme. À l'opposé, la ou les *variables dépendantes* sont les phénomènes mesurés qui devraient logiquement être influencés par la variable indépendante. Dans le cas du programme d'échange de seringues, la variable dépendante principale est le taux d'infection au VIH ou à l'hépatite.

En évaluation causale, il faut estimer le degré avec lequel une cause précise est responsable d'un effet mesuré. On le fait entre autres par élimi-

nation des causes alternatives possibles. Par exemple, si la table de billard est inclinée, le mouvement de la boule noire (l'effet) ne peut être entièrement attribué au mouvement de la boule blanche qui la percute (la cause). Bien qu'il soit possible que la boule blanche soit la cause de la trajectoire de la boule noire, elle n'en est pas nécessairement la cause, ni nécessairement la seule cause. Dans ce cas, l'attribution causale est ambiguë.

Pour minimiser l'ambiguïté des conclusions, il est généralement souhaitable de mesurer, outre la variable dépendante, certaines autres variables susceptibles d'affecter le résultat du programme et (ou) d'être affectées par ce résultat. Dans l'exemple du programme d'échange de seringues, en plus de mesurer les taux d'infection, on pourrait aussi mesurer un ensemble de variables connexes qui pourraient également influencer le taux d'infections rapporté, comme les attitudes envers les pratiques sécuritaires et les comportements des participants. Ainsi, en plus des résultats médicaux, des mesures des changements d'attitudes et (ou) de l'évolution de divers comportements à risque (tels que les relations sexuelles non protégées) pourraient aider à déterminer avec plus de précision les éléments qui ont contribué au succès ou à l'échec de l'intervention, mais aussi à identifier les éléments qui pourraient être modifiés pour l'améliorer. Dans ce cas précis, on a montré que le programme d'échange de seringues a contribué à augmenter de façon notable le pourcentage de participation aux programmes thérapeutiques visant le changement de comportement des toxicomanes, et aussi à réduire le taux de criminalité dans le voisinage.

Les devis de collecte de données en évaluation causale

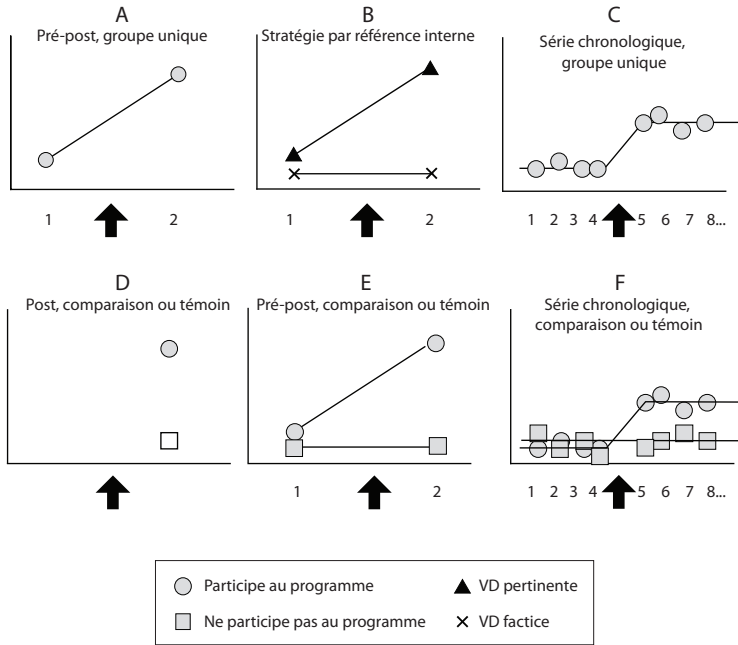
La détermination de la causalité implique des collectes de données organisées de manière précise par l'utilisation de devis. Il existe plusieurs types de devis qui permettent de faire des inférences causales, mais le degré de certitude des conclusions causales des différents types de devis varie. Ils se distinguent les uns des autres par leur capacité de réduire l'ambiguïté causale.

La figure 6.1 présente six devis de collecte de données utilisés dans la pratique évaluative. Bien qu'il en existe beaucoup d'autres (l'ouvrage original de Cook et Campbell, 1979, en fournit une liste exhaustive, actualisée en 2002 par Shadish, Cook et Campbell), nous centrons notre discussion sur ces six devis, car ce sont ceux qui sont le plus fréquemment utilisés dans la pratique, particulièrement pour les évaluations de programmes de grande envergure.

Les devis A, B et C sont des devis non expérimentaux, alors que les devis D, E et F sont expérimentaux ou quasi expérimentaux. Les devis A, B et C sont parfois nommés évaluations à groupe unique parce qu'ils permettent de comparer les résultats obtenus par des personnes après une intervention à ceux qui ont été obtenus par ces mêmes personnes avant l'intervention. Toutefois, alors que le devis A n'est pas en mesure d'offrir de conclusions causales qui soient valides, les devis B et C entraînent des inférences causales ambiguës. Les devis D, E et F sont quant à eux des devis d'évaluations expérimentales ou quasi expérimentales qui permettent de comparer les résultats obtenus par les personnes cibles à ceux ayant été obtenus par un (ou plusieurs) groupes indépendants de personnes qui n'ont pas bénéficié du traitement ou de l'intervention. En principe, ces trois derniers devis permettent tous de faire des attributions causales, mais la certitude des conclusions est plus forte pour les évaluations expérimentales que pour les évaluations quasi expérimentales. Nous reviendrons sur la distinction entre les devis expérimentaux et quasi expérimentaux après la présentation des six types de devis.

Dans chacun des graphiques, les informateurs qui participent au programme sont représentés par des cercles, et ceux qui n'y participent pas, par des carrés. Les mêmes mesures sont prises sur les deux groupes. L'abscisse des graphiques représente le moment de la collecte de l'information, alors que la flèche noire indique le moment où l'intervention a lieu. À titre d'exemple, pour le devis A, nous voyons que des mesures sont prises une fois avant et une fois après une intervention. Nous voyons aussi que les données ne sont recueillies que pour le groupe qui fait l'objet de l'intervention ou qui fait partie du programme devant être évalué. À l'opposé, dans le devis F, nous voyons que des données sont recueillies plusieurs fois avant et plusieurs fois après l'intervention, tant auprès du groupe qui bénéficie de l'intervention qu'auprès d'un autre groupe qui n'en bénéficie pas. Le devis B – la stratégie par référence interne (SRI) – est un cas spécial qui sera discuté séparément, car il repose sur une logique de collecte des informations qui diffère de celle qui est requise par les autres devis.

Le devis A (pré-post à groupe unique) comporte des mesures prises seulement auprès des participants au programme. La variable dépendante est mesurée à deux reprises, une fois avant et une fois après l'intervention. L'analyse statistique consiste alors à déterminer s'il existe une différence statistiquement significative entre les moments des deux collectes de

FIGURE 6.1
Six devis d'évaluation


données (à l'aide d'un test t paillé ou d'un test de McNemar¹) et à déterminer l'ampleur de cette différence. L'utilisation de ce devis pour évaluer un programme d'échange de seringues hypothétique exigerait que l'on détermine le taux moyen d'infections, tant pour l'année précédant l'introduction du programme que pour l'année suivant l'arrêt du programme, afin de vérifier si la différence entre ces deux taux est statistiquement significative. Ce devis permet de déterminer si les participants au programme sont « différents » après l'intervention, mais il ne permet pas d'identifier la cause de cette différence, puisque d'autres événements qui coïncident avec l'intervention pourraient aussi en être la cause.

1. À titre indicatif, nous donnons quelques exemples de tests statistiques adéquats pour chacun des devis. Il nous faut toutefois préciser que la liste des tests possibles pourrait à elle seule faire l'objet d'un autre chapitre, si nous prenions en considération les types de distribution, la taille de l'échantillon et la nature du problème évalué. Nous renvoyons donc le lecteur aux ouvrages indiqués dans la sélection bibliographique qui clôt ce chapitre.

Le devis C (série chronologique à groupe unique), tout comme le devis A, nécessite la collecte des mesures dépendantes avant et après l'intervention, et cela seulement auprès des participants au programme. Il se distingue du modèle pré-post à groupe unique par le fait qu'il répète la collecte des informations plusieurs fois avant et plusieurs fois après l'intervention. L'analyse statistique des données diffère de celle qui est requise par le devis A, car avec le devis C, il ne faut pas simplement comparer la moyenne des résultats postintervention à la moyenne des résultats pré-intervention. Il faut exécuter une analyse de tendance (généralement une régression polynomiale) démontrant la présence d'une cassure nette coïncidant avec l'intervention entre les résultats pré- et les résultats post, ce que les statisticiens nomment une relation quadratique. Ce devis est souvent utilisé lorsqu'il est possible d'avoir accès à des données objectives qui peuvent être recueillies sans déranger les sujets de l'intervention. Ce devis serait particulièrement utile pour l'évaluation du programme d'échange de seringues, car il serait facile de mesurer les taux d'infection sur plusieurs années en consultant les registres des hôpitaux de la ville. Mais tout comme pour le devis A, il est possible qu'une variable autre que l'intervention soit la cause de l'effet constaté. Toutefois, étant donné que nous recueillons plusieurs mesures pré- et plusieurs mesures post, et que le changement brusque coïncide avec le traitement, nous avons ici une plus grande certitude quant au fait que l'intervention en est la cause.

Le devis D (post avec comparaison indépendante) est le devis « expérimental causal » le plus simple. Nous lui attribuons cette étiquette, car la comparaison est centrée sur la différence entre un groupe qui bénéficie d'une intervention et un autre qui n'en bénéficie pas². La collecte de données n'est réalisée qu'une seule fois, après l'intervention. L'analyse statistique (un test t ou un test de proportions) porte sur la différence entre les résultats moyens obtenus par ces deux groupes sur la variable dépendante après l'intervention. Puisque nous ne mesurons la variable dépendante qu'en une occasion, il n'est pas possible de mesurer le changement chez les participants. Par conséquent, la validité de l'inférence causale et la certitude que nous avons par rapport à cette conclusion dépendront en grande partie du degré d'équivalence initiale des deux groupes.

2. Par opposition, les devis à groupe unique sont qualifiés de non expérimentaux, car ils n'impliquent pas l'utilisation d'un groupe témoin. Nous nous étendrons davantage sur cette distinction dans les pages suivantes.

Le devis E (pré-post avec comparaison indépendante) est identique au devis D, mais une collecte d'information est également faite avant le début de l'intervention. Comme il s'agit d'un devis expérimental, les données seront donc recueillies auprès de deux groupes indépendants, l'un qui fait l'objet de l'intervention et l'autre, qui n'en bénéficie pas. L'analyse statistique³ servira non seulement à démontrer qu'il existe une différence pré-post, mais aussi que cette différence n'est constatée que pour le groupe qui aura participé à l'intervention. Il s'agit donc de démontrer l'existence d'une interaction statistiquement significative entre les conditions de traitement (groupe participe ou non à l'intervention) et le moment de la collecte d'information (pré-post).

Le devis F (série chronologique avec comparaison indépendante) est identique au devis C, mais il ajoute des collectes de données auprès d'un groupe qui n'aura pas bénéficié de l'intervention. La démonstration statistique consistera à établir la présence (ou l'absence) d'une relation quadratique seulement chez le groupe qui bénéficie de l'intervention, et cela à l'aide d'une régression polynomiale.

Le devis expérimental versus le devis non expérimental

En analysant les devis décrits à la figure 6.1, on remarque que c'est la présence d'un groupe de comparaison qui distingue les modèles expérimentaux et quasi expérimentaux des modèles non expérimentaux. Avec les devis non expérimentaux, on compare les résultats obtenus après l'intervention à ceux qui ont été obtenus avant l'intervention pour un même groupe de personnes. Par conséquent, il est tout à fait plausible que des variables connexes à l'intervention soient responsables d'une éventuelle différence pré-post. À strictement parler, il est donc inapproprié d'attribuer les différences pré-post (causales) au seul effet de la variable indépendante lorsqu'on utilise un devis non expérimental.

En revanche, la logique causale des modèles expérimentaux est facile à comprendre. En présumant que les deux groupes sont initialement

3. Parmi les méthodes les plus fréquemment utilisées pour l'analyse du devis E, on suggère depuis quelques années d'utiliser de préférence les trois suivantes : 1- la régression hiérarchique utilisant la mesure post-test comme variable dépendante, le score au prétest en tant que variable contrôle et le traitement en tant que variable indépendante ; 2- l'analyse de covariance (où la mesure prétest est la covariable) ; et 3- l'analyse de variance à mesures répétées (voir entre autres Gliner *et al.* [2003]).

identiques quant à la variable dépendante et que l'intervention n'est menée qu'avec un des groupes, la différence entre les deux ne peut être attribuée qu'à l'effet de l'intervention. Nous concluons donc que l'intervention est bien la cause de la différence. Naturellement, la validité de cette inférence repose sur l'équivalence initiale entre les groupes. Si les deux groupes ne sont pas identiques au départ, l'inférence causale est menacée, car la différence entre les groupes pourrait être attribuable à la différence initiale plutôt qu'au traitement.

Devis expérimental ou devis quasi expérimental ?

Il existe deux techniques générales pour constituer un groupe de comparaison. Dans un cas, nous pourrions commencer avec une seule cohorte de personnes qui seraient réparties aléatoirement en deux groupes. L'intervention serait effectuée sur un seul des groupes et l'autre deviendrait le groupe de comparaison qui, dans ce cas, serait qualifié de *groupe témoin*. Les devis qui exigent de constituer des groupes en faisant appel à la répartition aléatoire prennent le nom de *devis expérimentaux*. La répartition aléatoire est la technique qui assure le mieux l'équivalence des groupes.

Par ailleurs, pour des raisons pratiques ou éthiques, il n'est pas toujours possible de constituer un groupe témoin⁴. Par exemple, pour l'évaluation d'un programme d'échange de seringues, où la participation au programme serait faite sur une base volontaire, nous ne pourrions décider de refuser l'accès aux seringues à certains héroïnomanes choisis aléatoirement pour faire partie du groupe de comparaison. De la même façon, lors de l'évaluation d'un programme de distribution de moustiquaires visant à réduire l'incidence de la malaria, il ne serait pas éthique d'offrir des moustiquaires à seulement un échantillon de villageois. Enfin, les règles de l'aléa ne fonctionnent que lorsque le nombre de personnes dans chaque groupe est relativement important. Le nombre minimal de sujets par groupe étant relatif à la variabilité au sein de la population, ce nombre

4. Une exception provient des domaines médical et pharmaceutique, où la répartition aléatoire est presque toujours utilisée lorsqu'il s'agit d'évaluer l'efficacité de procédures et (ou) de médicaments. De plus, et particulièrement dans ce domaine de recherche évaluative, il est préférable de faire appel à trois groupes composés de façon aléatoire : un groupe reçoit le médicament expérimental, un groupe reçoit un placebo et un groupe ne reçoit ni l'un ni l'autre.

pourrait être relativement faible ($N \approx 30$ pour chaque groupe). Il devrait toutefois être plus important lorsque la variance est plus grande et (ou) lorsque la variable sous-jacente n'est pas distribuée normalement (voir Haccoun et Cousineau, 2010).

Lorsqu'il s'agit d'évaluer les impacts de programmes qui s'adressent à un petit nombre de personnes, la répartition aléatoire ne garantit pas nécessairement l'équivalence des groupes. Par conséquent, il est généralement plus facile de comparer le groupe expérimental à un groupe naturel qui lui ressemble. Par exemple, le groupe de comparaison pourrait être les héroïnomanes d'une autre ville canadienne de taille similaire, mais qui n'offre pas de programme d'échange de seringues. Ainsi, lorsque les groupes ne sont pas répartis aléatoirement, le devis devient un *devis quasi expérimental* et le groupe témoin prend le nom de *groupe de comparaison*. Les modèles quasi expérimentaux permettent de faire des inférences causales, mais celles-ci ne sont jamais aussi certaines que celles des devis expérimentaux. En effet, avec les devis quasi expérimentaux, il reste toujours un doute sur l'équivalence initiale des groupes. La différence observée pourrait donc être attribuable à l'inégalité initiale tout autant qu'au traitement. Par exemple, pour l'évaluation du programme d'échange de seringues ou de celui qui portait sur les moustiquaires, le problème reste : est-ce que la population d'héroïnomanes dans les deux villes ou la densité de moustiques dans les deux villages est véritablement similaire ?

Somme toute, la validité des inférences causales dépend du degré de certitude que nous avons quant au fait que la différence observée puisse être attribuée à la cause présumée. Ainsi, nous voyons que l'inégalité initiale des groupes constitue une menace importante à la validité des inférences causales, mais il ne s'agit toutefois pas de la seule.

L'inférence causale et les concepts de validité interne et externe

La notion de démonstration causale implique nécessairement l'idée de preuve qu'un effet observé ne peut pas avoir été créé par autre chose que la variable causale en jeu. En réalité, il faut reconnaître qu'aucune technique ou procédure ne peut entièrement garantir cette conclusion. Il s'agit plutôt d'une probabilité qui peut être plus ou moins forte.

Dans un des ouvrages classiques les plus importants de ce domaine, Campbell et Stanley (1963) ont analysé les risques relatifs à la validité

interne et externe associés à l'utilisation de divers devis de recherche. La validité interne est liée aux possibilités d'interprétation des résultats, alors que la validité externe porte sur la généralisation à la population des résultats obtenus dans une évaluation. L'encadré suivant énumère les menaces qui, selon ces auteurs, pèsent sur la validité interne et externe. Chacun de ces obstacles est ici illustré dans le cadre d'un programme de formation portant sur la réduction du stress au travail.

Les obstacles à la validité : l'exemple d'un programme de réduction du stress au travail

Obstacles à la validité interne

1. *Histoire* : Les événements qui coïncident avec le traitement pourraient être à l'origine des résultats obtenus.
Au moment de la formation, vous avez vu un documentaire qui traite du même sujet. Un changement du niveau de stress pourrait être causé par la formation, par le documentaire ou par les deux.

2. *Maturation* : L'évolution naturelle des participants affecte les effets répertoriés.
Le simple passage du temps fait en sorte que votre réaction au stress change.

3. *Mesure* : La mesure faite avant le traitement affecte la réaction de la personne après le traitement.
La mesure du stress avant la formation vous fait découvrir que vous êtes une personne stressée.

4. *Instrumentation* : Un changement dans la mesure elle-même.
Les questionnaires utilisés pour mesurer votre niveau de stress avant et après la formation ne sont pas les mêmes.

5. *Régression statistique* : Lorsque l'assignation au programme est déterminée par les résultats au prétest, leurs résultats au post-test régressent vers la moyenne.
On mesure le niveau de stress par questionnaire avant la formation. Seules les personnes ayant des résultats élevés reçoivent la formation. Lors du post-test, les scores obtenus sur la mesure de stress auront tendance à s'abaisser pour ces personnes.

6. *Sélection* : Les participants qui sont choisis pour faire partie du programme ne sont pas équivalents à ceux qui n'en font pas partie.
La formation est offerte aux seules personnes qui désirent la suivre alors que le groupe de comparaison est composé de celles qui ne le désirent pas. La réaction au stress est différente pour celles qui s'inscrivent ou qui ne s'inscrivent pas. Le degré de motivation à réduire le stress diffère selon que les personnes s'inscrivent ou non au programme, ce qui pourrait expliquer la différence de niveau de stress ressenti à la suite de la participation au programme.

7. *Mortalité expérimentale* : La perte de sujets qui faisaient partie du programme ou du groupe de comparaison n'est pas la même.
Les personnes qui sont très (ou très peu) stressées quittent le programme avant sa fin et, par conséquent, il n'est plus possible de mesurer leur niveau de stress post-formation.

8. *Interaction entre sélection et maturation* : l'évolution naturelle des personnes dans le groupe de comparaison diffère de l'évolution naturelle de celles qui participent au programme.

Obstacles à la validité externe

1. *Réactivité au prétest* : Le simple fait d'être mesuré affecte les mesures subséquentes.
En mesurant l'état initial de stress ressenti par le participant, celui-ci réalise qu'il est stressé.

2. *Interaction entre la sélection des participants et le programme* : Le traitement administré interagit avec la sélection différentielle.
Les personnes choisies pour le groupe de traitement sont plus (ou moins) sensibles au stress que celles qui sont choisies pour le groupe de comparaison.

3. *L'effet Hawthorne*: Le fait de se savoir « étudié » affecte les résultats des participants. Les participants à l'étude réalisent qu'on leur prête attention, ce qui altère leur comportement.

4. *Les effets résiduels des traitements multiples*: Les effets d'un premier traitement contaminent les effets produits par les traitements subséquents.

Si les sujets qui participent au programme étaient préalablement traités par des médicaments, l'effet du programme de formation sur la réduction du stress serait indéterminé.

Source : Adapté de Campbell et Stanley, 1963, p. 5-6.

La liste produite par Campbell et Stanley est importante, car elle propose une manière pratique d'évaluer la qualité de divers devis pour l'évaluation. Les devis qui produisent les inférences causales les moins ambiguës sont ceux qui éliminent le plus ces menaces à la validité interne et externe.

L'évaluation de la validité des divers devis de collecte de données

Les tableaux 6.1 et 6.2 reprennent les devis décrits à la figure 6.1 et les analyses sur les risques relatifs qui pèsent sur leur validité interne et externe. Ces évaluations sont tirées de Campbell et Stanley (1963).

TABLEAU 6.1

Les risques aux validités interne et externe pour les devis d'évaluation A et C

	A Pré-post, groupe unique	C Série chronologique, groupe unique
Validité interne		
1. Histoire	E	E
2. Maturation	E	F
3. Testage	E	F
4. Instrumentation	E	M
5. Régression statistique	M	F
6. Sélection différentielle	F	F
7. Mortalité expérimentale	F	F
8. Interaction sélection / maturation	E	F
Validité externe		
1. Réactivité au prétest	E	E (F si mesures objectives)
2. Interaction sélection / programme	E	M
3. Effet Hawthorne	E	M
4. Effets résiduels de traitements multiples	X	X
E : risque élevé; F : risque faible; M : risque moyen; X : ne s'applique pas		

En comparant le nombre de risques E (élevés) et F (faibles) pour les deux modèles, nous voyons clairement que le modèle A (pré-post à groupe unique) est nettement inférieur au modèle C. En fait, le modèle C peut éliminer la plupart des risques relatifs à la validité des conclusions, alors que le modèle A présente tous les risques, sauf ceux qui se rapportent à la non-équivalence des groupes traité et de comparaison. Le risque 9 (réactivité au prétest) est noté E (élevé) pour le devis par série chronologique, mais il devient faible (F) lorsque les mesures recueillies sont des données objectives qui peuvent être mesurées sans que les personnes puissent en être conscientes.

TABLEAU 6.2
Les risques aux validités interne et externe
pour les devis d'évaluation D, E et F

	D Post, groupe témoin ou de comparaison	E Pré-post, groupe témoin ou de comparaison	F Série chronolo- gique, groupe témoin ou de comparaison
Validité interne			
1. Histoire	F	F (F)	F
2. Maturation	F (M)	M (F)	F
3. Testage	F	F (F)	F
4. Instrumentation	F	F (F)	F
5. Régression statistique	F	F (F)	F
6. Sélection différentielle	F (E)	E (F)	F
7. Mortalité expérimentale	F (E)	E (F)	F
8. Interaction sélection / maturation	F (E)	E (F)	F
Validité externe			
1. Réactivité au prétest	X	E	E (F si mesures objectives)
2. Interaction sélection / programme	M (E)	M	E
3. Effet Hawthorne	M (X)	M	F
4. Effets résiduels de traitements multiples	X	X	X

E: risque élevé; **F**: risque faible; **M**: risque moyen; **X**: ne s'applique pas
Les risques sont évalués pour les devis qui se servent d'un groupe témoin. Lorsque la comparaison est faite avec un groupe de comparaison, le risque est indiqué entre parenthèses.

Un cas particulier : le devis B (stratégie par référence interne)

En pratique, particulièrement dans les programmes de grande envergure, il est très difficile de constituer des groupes de comparaison, et encore plus difficile d'établir des groupes témoins. Dans ces cas, on préfère employer le devis par série chronologique, bien que son utilisation ne soit pas toujours possible pour des raisons pratiques (par exemple, il n'est pas toujours possible d'avoir accès à des données objectives et l'utilisation de mesures répétées peut créer des problèmes de validité interne, particulièrement les effets dus au testage et à l'instrumentation). Pour cette raison, la plupart des évaluations faites dans la pratique (comme pour l'évaluation des programmes de formation – voir Saks et Haccoun, 2007) se limitent au devis A : le devis pré-post à groupe unique. Cependant, comme nous l'avons vu au tableau 6.1, l'utilisation du devis A ne permet pas de faire des inférences causales adéquates.

Il est cependant possible d'améliorer le devis A en faisant appel au devis B, la stratégie par référence interne conçue par Haccoun et Hamtiaux (1994). Cette technique propose d'inclure dans les mesures avant et après non seulement des mesures qui devraient être affectées par le programme (les mesures « pertinentes » à la figure 6.1), mais aussi des variables connexes à l'intervention, c'est-à-dire qui ne sont pas la cible de l'intervention (les mesures « factices »). Si le changement pré-post évalué à l'aide des mesures pertinentes est plus grand que le changement évalué avec les mesures factices, nous obtenons une preuve causale de l'efficacité du programme. À l'opposé, si les changements dans les variables factices et pertinentes sont statistiquement égaux, il n'est pas possible de conclure à l'efficacité du programme.

La forme statistique requise est une analyse de variance (avec mesures répétées) où une interaction statistique entre les mesures (pertinentes-factices) et le moment de mesure (pré-post) signifiera que le programme est efficace. Lorsque les améliorations sont équivalentes pour les deux séries de mesures, nous ne pouvons pas conclure que le programme est efficace. En langage statistique, on dira que la stratégie par référence interne est susceptible d'entraîner des erreurs de type II (Haccoun et Cousineau, 2007), erreurs que nous expliquerons ici brièvement.

Lorsque les deux séries de variables montrent des améliorations équivalentes, il n'est pas possible de choisir entre deux interprétations potentielles : soit a) le programme n'est pas efficace et les changements sont

attribuables à l'un ou l'autre des risques liés aux validités, ou b) le programme est efficace et les participants généralisent les impacts du programme aux mesures factices. Par exemple, pour le programme d'échange de seringues, en plus de mesurer les taux d'infection par VIH et hépatite (les variables pertinentes), nous pourrions recueillir des données sur les comportements sexuels non sécuritaires (variables factices). Mais la difficulté, dans un cas semblable, est que le programme d'échange pourrait avoir conscientisé les héroïnomanes au problème des comportements non sécuritaires, occasionnant ainsi des changements de leurs comportements sexuels, mais aussi de leurs comportements strictement liés à l'utilisation de l'héroïne. Dans ce cas, l'intervention est efficace dans les faits, mais le devis B ne peut en établir la preuve.

Des vérifications empiriques (Haccoun et Hamtiaux, 1994; Frese *et al.*, 2003) démontrent que la stratégie par référence interne (SRI) peut offrir des inférences causales équivalentes à celles qui sont produites par le devis expérimental E. Le devis SRI est donc supérieur au modèle A, mais il doit tout de même être considéré comme un modèle causal minimal à n'être utilisé que lorsqu'il n'est pas possible de faire appel à un devis autre que celui du modèle A. Il ne doit jamais être substitué aux autres devis (D à F) si leur emploi est possible.

L'analyse des données : signification statistique et signification pratique

Les comparaisons requises dans le domaine de l'évaluation exigent l'utilisation de procédures statistiques, principalement des tests comparatifs comme les tests t et (ou) les analyses de variance/covariance ou des régressions linéaires. Lorsque ces tests font ressortir des différences (ou des corrélations) « statistiquement significatives », nous concluons (en tenant compte d'un risque d'erreur préétabli, par exemple 5 %) qu'il existe une différence « réelle », c'est-à-dire une différence qui apparaîtra aussi dans l'ensemble de la population. Nous pourrions ainsi conclure que l'incidence du VIH et de l'hépatite est différente *dans la population* depuis l'instauration du programme d'échange de seringues. Toutefois, selon le nombre de participants et certaines autres caractéristiques propres à l'étude⁵, il

5. Voir Haccoun et Cousineau (2007) pour une discussion plus approfondie des caractéristiques qui affectent les chances de conclure à la signification statistique.

est possible que même des différences triviales puissent être statistiquement significatives. Aussi serait-il erroné de conclure que l'impact du programme est substantiel en ne nous basant que sur la simple signification statistique. De fait, *la signification statistique est une preuve essentielle, mais minimale et insuffisante*, de l'efficacité d'un programme.

Il est donc important de soumettre les données à une analyse additionnelle, soit la détermination de l'ampleur de la différence; c'est ce que les statisticiens nomment *la taille de l'effet*. Ces procédures statistiques additionnelles, telles que le coefficient de détermination r_{xy}^2 ou le ratio de corrélation η^2 (éta au carré), permettent de répondre à la question suivante: quelle proportion de toutes les différences relevées par le test comparatif peut être expliquée par le traitement ou le programme? Si la différence est grande, cela indiquera que le programme pourrait avoir un impact pratique d'envergure: il aurait alors un niveau élevé de signification pratique. Mais si la différence est petite, il y aurait peut-être moins de raisons de poursuivre le programme, particulièrement lorsque celui-ci est coûteux ou difficile à mettre en œuvre. Sa signification pratique serait faible.

En résumé, les évaluations doivent d'abord tester la signification statistique et, par la suite, démontrer la taille de l'effet. Advenant le cas où la différence serait statistiquement significative, l'interprétation du succès ou de l'échec d'un programme serait alors basée principalement sur la taille de l'effet, car celle-ci indique la portée réelle – la signification pratique – du programme.

Quelques considérations sur le choix d'un modèle d'évaluation causale

L'objectif idéalisé d'une évaluation causale est de produire une attribution claire de la causalité d'un effet. Par conséquent, l'évaluateur choisira, parmi les modèles d'évaluation disponibles, celui qui offrira la conclusion causale la plus solide. Mais l'évaluateur se doit aussi d'être sensible à l'éthique de sa démarche ainsi qu'à l'aspect pratique et aux coûts associés à ce choix: les modèles ne sont pas tous également ou facilement applicables et appropriés à tous les contextes (voir chapitre 9).

Dans la pratique, les modèles strictement expérimentaux (avec groupe témoin ou groupe de comparaison) sont généralement difficiles à utiliser. Le devis par série chronologique est sans doute celui qui reste le plus commode. Mais là encore, il faut prendre en considération le fait que

celui-ci peut être efficacement employé surtout lorsque l'on a accès à des données objectives pouvant être recueillies sans créer de réactivité chez les participants. En l'absence de ces conditions, et lorsque nous devons nous limiter à un modèle pré-post à groupe unique, il sera préférable de l'améliorer en faisant appel à la stratégie par référence interne.

En bout de ligne, le choix d'un modèle dépend du risque associé à une erreur dans l'attribution de la causalité. Si le fait de tirer une conclusion fautive n'a qu'une importance secondaire, alors on peut employer un modèle d'évaluation plus simple comme le modèle B. À l'inverse, lorsque les conséquences d'une telle erreur pourraient être graves, seuls les modèles d'évaluations complexes – les modèles expérimentaux – doivent être retenus.

Pour aller plus loin :

COLLINS, L. M. et J. L. HORN (1991), *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions*, Washington, American Psychological Association.

L'ouvrage aborde certains aspects de la mesure du changement, à l'aide, entre autres, des modèles d'équations structurales et des variables latentes.

GLINER, J. A., G. A. MORGAN et R. J. HARMON (2003), « Pretest-posttest comparison group designs: Analysis and interpretation », *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 42, n° 4, p. 500-503.

Ce court article de vulgarisation présente les forces et faiblesses des analyses basées sur la régression hiérarchique, l'analyse de covariance, l'analyse de variance avec mesures répétées et l'approche du score de différence.

HACCOUN, R. R. et D. COUSINEAU (2007), *Statistiques: Concepts et applications*, Montréal, Les Presses de l'Université de Montréal.

Introduction à la logique et aux techniques d'inférence. La présentation est généralement non mathématique et axée sur les concepts et les interprétations qui peuvent découler des diverses techniques statistiques. Les pages Web associées au livre présentent des problèmes et des banques de données qui peuvent être analysées à l'aide de SPSS, un logiciel professionnel d'analyse statistique.

SAKS, M. A. et R. R. HACCOUN (2007), *Managing Performance Through Training and Development*, 4^e édition, Toronto, Nelson Publishers.

Bien qu'il traite principalement de la formation en milieu de travail, cet ouvrage comprend deux chapitres complets consacrés à l'évaluation, à ses mesures, à ses techniques et à ses devis.

7

Les méthodes mixtes

Pierre Pluye

Les méthodes mixtes de recherche et d'évaluation combinent les méthodes de collecte et d'analyse de données qualitatives et quantitatives. Elles sont de plus en plus fréquemment utilisées et, au fil des ans, plusieurs termes ont été proposés pour les nommer (par exemple, multiméthodes). Johnson (2007) a effectué une synthèse critique de la littérature scientifique sur les méthodes mixtes et a proposé une définition de ces méthodes pour la recherche. Cette recension des écrits suggère de définir les méthodes mixtes pour l'évaluation de programmes de la manière suivante. Une évaluation mixte est un type d'évaluation dans laquelle un expert ou une équipe d'experts combinent les approches ou devis méthodologiques avec les techniques de collecte et d'analyse de données qualitatives et quantitatives pour approfondir la compréhension, ou interpréter, ou corroborer des résultats d'évaluation.

Les trois principales raisons de combiner méthodes qualitatives et quantitatives sont les suivantes. Les évaluateurs peuvent avoir besoin des méthodes qualitatives pour interpréter des résultats quantitatifs. Ils peuvent avoir besoin des méthodes quantitatives pour généraliser des résultats qualitatifs. Enfin, ils peuvent avoir besoin d'explorer ou de mieux comprendre un phénomène nouveau (méthodes qualitatives) et d'en mesurer l'ampleur, l'évolution, les causes et les effets (méthodes quantitatives).

Nous remercions Lucie Nadeau, Marie-Pierre Gagnon, Roland Grad, Janique Johnson-Lafleur et Frances Griffiths pour leur contribution à une version précédente de ce chapitre (2009).

Dans le tableau 7.1, nous présentons une liste non exhaustive de méthodes qualitatives et quantitatives susceptibles d'être combinées sous forme de méthodes mixtes. Les lecteurs qui ne les connaissent pas bien peuvent se référer aux dictionnaires de Muchielli (2004) pour les méthodes qualitatives et de Porta (2008) pour les méthodes quantitatives.

TABLEAU 7.1**Exemples de méthodes qualitatives et quantitatives**

Méthodes qualitatives	Méthodes quantitatives
Communication et information - Analyse de discours Idéaux-types et « patterns » - Analyse des événements - Théorisation ancrée - Étude de cas qualitative Actions sociales - Herméneutique - Histoires de vie - Phénoménologie Société et interactions sociales - Ethnographie - Ethnométhodologie	Études randomisées - Essai contrôlé randomisé Études non randomisées - Essai contrôlé - Étude de cohorte - Étude de comparaison de cas - Étude pré-post / séries temporelles - Enquête transversale (analytique) Études descriptives - Étude de prévalence ou d'incidence (sans groupe de comparaison) - Série de cas - Rapport de cas

À titre d'illustration, une méthode mixte d'évaluation des services de santé mentale en milieu multiethnique pourrait combiner une enquête quantitative transversale (questionnaire structuré) avec une étude ethnographique (documents, entrevues et observations), permettant ainsi de mieux comprendre les résultats de l'enquête et d'élaborer un nouveau programme pour faciliter l'accès aux services (Groleau *et al.*, 2007). Dans cet exemple, l'enquête seule aurait suffi pour révéler les problèmes d'accessibilité, mais elle aurait été insuffisante pour permettre de créer un programme approprié. Comme l'indiquent Sharp et Frechtling (1997), « en se concentrant principalement sur des techniques quantitatives, les experts peuvent passer à côté de certaines parties importantes de l'histoire. Les évaluateurs expérimentés ont maintes fois constaté que de meilleurs résultats sont obtenus en utilisant des méthodes mixtes d'évaluation » (p. 2). De fait, on a depuis longtemps combiné ces méthodes en recherche et en évaluation, et cette combinaison peut même être vue comme inévitable dans le domaine de l'évaluation (Rallis et Rossman, 2003, p. 491). Mais ce n'est que récemment que ce type de combinaisons

a été conceptualisé en termes de méthodes mixtes (Creswell et Plano Clark, 2010 ; Greene, 2006 ; Johnson, 2007 ; Tashakkori et Teddlie, 2010).

En général, les méthodes mixtes sont utilisées en vue de combiner les forces et compenser les limites respectives des méthodes quantitatives et qualitatives. D'une part, les évaluations quantitatives examinent habituellement des associations de facteurs (par exemple, des causes et leurs effets) pouvant être généralisées à toute une population, notamment grâce à des inférences statistiques. D'autre part, les évaluations qualitatives fournissent des descriptions détaillées de phénomènes complexes fondées empiriquement sur un contexte spécifique pouvant être théoriquement applicables dans d'autres situations (par exemple, une étude ethnographique). En particulier, la partie qualitative d'une évaluation mixte peut fournir une meilleure compréhension du développement des programmes complexes (planification, implantation et pérennisation), et une compréhension approfondie des raisons pour lesquelles ces programmes fonctionnent bien dans certains contextes (ou ne fonctionnent pas dans d'autres situations).

Ce chapitre vise à présenter notre conception des méthodes mixtes, puis à décrire les approches, devis et techniques de collecte et d'analyse de données qualitatives et quantitatives qui sont utiles pour planifier ou implanter des méthodes mixtes pour l'évaluation des programmes. Une illustration sera tirée de notre travail en évaluation des technologies. Dans les sections finales, nous proposerons des critères de qualité pour planifier ou évaluer les méthodes mixtes, et pour effectuer une recension des écrits mixte pour l'évaluation des programmes.

Les méthodes mixtes et les débats entre méthodes quantitatives et qualitatives

La méthodologie des méthodes mixtes peut être conceptualisée à partir des débats méthodologiques entre les tenants des méthodes quantitatives et ceux des méthodes qualitatives. Ces débats suggèrent deux idéaux-types méthodologiques, souvent considérés comme deux paradigmes concurrents, qui sont respectivement associés aux méthodes quantitatives et qualitatives : l'empirisme logique (matérialisme, réalisme et arguments objectifs habituellement associés aux méthodes quantitatives) et le constructivisme (idéisme, relativisme et arguments subjectifs habituellement associés aux méthodes qualitatives). Un paradigme est défini comme étant une norme (ou un

standard) institutionnelle des sciences. Bien que selon Kuhn (1983), l'évolution des sciences soit caractérisée par des transitions successives d'un paradigme à l'autre, des révolutions, de nombreux historiens ou philosophes des sciences observent que plusieurs paradigmes ou « visions du monde » coexistent (Morgan, 2007). Par exemple, l'empirisme logique, le paradigme dominant des disciplines scientifiques au xx^e siècle, est contesté par les tenants du constructivisme (Denzin et Lincoln, 1994), et coexiste avec le constructivisme. Dans cette section, nous allons brièvement présenter ces deux idéaux-types, leurs principales forces et faiblesses en évaluation de programmes, et la façon dont ces idéaux-types permettent de conceptualiser les méthodes mixtes. Le terme « idéal », dans le concept (outil d'analyse) de l'idéal-type proposé par Max Weber, désigne l'idée, et un idéal-type est le groupement d'un certain nombre de caractéristiques typiques des phénomènes sociaux qui a pour but d'aider à la compréhension de ces derniers.

L'empirisme logique vise à étudier des phénomènes selon des lois empiriques, par exemple la force de l'association entre des facteurs importants de l'implantation des programmes (mesurés en valeurs de variables indépendantes) et les résultats observés en termes d'implantation (mesurés en valeurs de variables dépendantes). Les relations causales constituent un exemple de ces lois empiriques et sont présentées en détail au chapitre 6, intitulé « Attribution et causalité des effets d'un programme ». L'empirisme logique dérive du positivisme proposé au xix^e siècle, intègre les critiques postpositivistes du xx^e siècle et propose habituellement (mais pas nécessairement) que les lois empiriques soient informées par une théorie qui préexiste à la démarche de recherche (approche déductive ou confirmatoire). Par exemple, les empiristes logiques assument habituellement (mais encore là, pas nécessairement) qu'il n'y a pas de causalité sans organisation normative préalable du champ d'exploration empirique par une explication scientifique logique. L'étude de la causalité est alors fondée sur des hypothèses formulées à l'aide de relations conditionnelles (langage de la science selon l'empirisme logique)¹.

En évaluation de programmes, des biais peuvent influencer la validité des études quantitatives expérimentales ou quasi expérimentales. Les

1. Nous remercions François Béland, de l'Université de Montréal, pour ses commentaires sur l'empirisme logique. Ce paragraphe ne rend évidemment pas compte de tous les travaux des empiristes logiques, et notamment de ceux qui ont développé une logique « inductive ».

sources des biais sont plus ou moins contrôlées selon les devis utilisés, les études randomisées étant considérées comme les plus valides pour examiner des relations causales (plus valides que les études non randomisées, par exemple). Les études randomisées ont néanmoins leurs propres faiblesses (Mays *et al.*, 2005 ; Pawson, 2005). Ainsi, une évaluation avec essai randomisé contrôlé peut être inadéquate selon le contexte ou la question, et elle peut être contredite par des études non randomisées. Par exemple, les études randomisées sont moins appropriées que les études de cohorte pour évaluer les rares effets négatifs à long terme des programmes (la randomisation de milliers de personnes peut être nécessaire pour détecter ces effets, mais elle risque d'être irréalisable ou trop coûteuse). De plus, les programmes peuvent être conceptualisés en tant qu'interventions sociales complexes agissant sur des systèmes sociaux complexes. Or, les études randomisées fournissent habituellement peu ou pas de résultats empiriques expliquant pourquoi les effets des interventions varient selon différents contextes d'implantation.

Quant au constructivisme, il vise à découvrir et à explorer des phénomènes complexes, à décrire leurs variations selon leur contexte et à en proposer des interprétations différentes pour mieux les comprendre (approche inductive ou exploratoire). Le constructivisme dérive des approches dites compréhensives développées en Allemagne dans les années 1920 par Weber, et à Chicago, par les fondateurs d'une approche sociologique américaine appelée « interactionnisme symbolique ». Les chercheurs qualitatifs soulignent la nature socialement construite de la réalité, la relation intime entre le chercheur, le phénomène étudié et leur contexte (Denzin et Lincoln, 1994). Ils examinent les aspects historiques, interprétatifs et narratifs en fonction de leur contexte de production culturel, économique et social (Abbott, 1998).

Les méthodes qualitatives ont été développées de manière concomitante dans plusieurs disciplines et utilisent différentes procédures interprétatives (par exemple, l'ethnographie en anthropologie et les ethnométhodes en sociologie). Les méthodes qualitatives ont du succès, en évaluation de programmes, car elles satisfont les besoins des décideurs et des professionnels en termes de compréhension riche et détaillée des interventions sociales complexes (Patton, 2002), compréhension utile pour la planification, l'implantation et la pérennisation des programmes. Plus particulièrement, les études qualitatives fournissent des résultats empiriques expliquant pourquoi

les effets des interventions varient selon les différents contextes d'implantation. Cependant, les méthodes qualitatives sont critiquées parce que les résultats peuvent être idiosyncrasiques à un contexte spécifique (aucune connaissance transférable à un autre contexte), ou fondés sur des procédures interprétatives tacites, non reproductibles et non vérifiables (Huberman et Miles, 1991).

Les méthodes mixtes questionnent ces idéaux-types, et suggèrent des types méthodologiques différents. Par exemple, les études quantitatives exploratoires ne correspondent pas à l'idéal-type méthodologique de l'empirisme logique. Les résultats statistiques ou épidémiologiques peuvent être considérés comme des constructions sociales utiles pour explorer et comprendre des phénomènes sociaux (constructionnisme ou réalisme constructif). Ainsi, Kallerud et Ramberg (2002) favorisent l'interprétation constructiviste des enquêtes. Autre exemple, les études qualitatives inductives-déductives ne correspondent pas à l'idéal-type méthodologique du constructivisme. Les résultats qualitatifs sont souvent intégrés dans des recherches examinant des relations causales, et sont considérés comme des arguments scientifiques valides pour générer des théories ou des hypothèses (post-positivisme). Ainsi, selon Campbell (1988), une considération qualitative des différents contextes de mesure permet de proposer des hypothèses rivales plausibles.

Le débat entre méthodes quantitatives (ci-après QUANT) et méthodes qualitatives (ci-après QUAL) permet de concevoir les méthodes mixtes (ci-après MM) par rapport aux deux idéaux-types méthodologiques résumés ci-dessus. Tel que précisé en introduction de ce chapitre, en évaluation de programmes, les méthodes quantitatives et qualitatives coexistent, et cette coexistence contribue au développement des MM et, par ces dernières, au développement de combinaisons d'approches, de devis et de techniques.

Les méthodes mixtes : approches, devis et techniques de collecte et d'analyse de données

La présente section vise à présenter toutes les possibilités de combinaisons et d'intégration des approches, des devis et des techniques pour faciliter la planification et l'évaluation des MM. Tout comme Johnson (2007), nous sommes d'avis qu'il est légitime de considérer comme MM toute combi-

raison d'approches méthodologiques, de devis et de techniques de collecte ou d'analyse de données, mais uniquement lorsque ces combinaisons satisfont trois conditions: 1) au moins une méthode QUAL et une méthode QUANT sont combinées (par exemple, l'ethnographie et l'expérimentation); 2) chaque méthode est utilisée de façon rigoureuse par rapport aux critères généralement admis (voir la section suivante « Planification et qualité des MM en évaluation de programmes »); et 3) la combinaison des méthodes se traduit par une intégration des approches méthodologiques, des devis ou des techniques (par exemple, des observations et des mesures), et des données et (ou) des résultats. Inversement, selon nous, ne sont pas des MM: 1) une méthode QUANT avec collecte ou analyse de données qualitatives qui ne réfère pas à une méthode QUAL (par exemple, une étude de prévalence utilisant un questionnaire structuré auto-administré avec principalement des questions à réponse fermée, et quelques questions requérant des commentaires libres), ou vice-versa; 2) les combinaisons de méthodes QUANT; 3) les combinaisons de méthodes QUAL; et 4) les juxtapositions de méthodes QUAL et QUANT sans aucune intégration d'approches, de devis, de techniques, de données et (ou) résultats QUAL et QUANT.

Les approches

Les experts des MM comme Bryman (sociologie) et Cresswell (psychologie), ainsi que Greene, Morgan, Tashakkori et Teddlie (éducation) proposent différentes typologies des approches en MM. Nous nous inspirons ici des propositions de Teddlie et Tashakkori (2003). Évidemment, les MM sont impossibles lorsque les méthodes QUANT et QUAL sont considérées comme incompatibles ou *incommensurables* (positivisme radical ou constructivisme radical). Trois approches principales soutiennent les MM en évaluation de programmes.

L'approche *complémentaire* indique que les résultats QUAL et QUANT sont présentés séparément, mais qu'un élément QUAL contribue à un élément QUANT, ou vice-versa. Par exemple, les résultats QUAL et QUANT sont clairement distincts dans la section « Résultats » des rapports d'évaluation ou des publications, mais au moins une phrase suggère que les résultats QUAL et QUANT sont complémentaires.

L'approche par *tension dialectique* souligne les divergences toujours possibles entre les résultats QUAL et QUANT. Par exemple, des divergences

entre les résultats QUANT et QUAL sont clairement énoncées dans la section « Résultats » des rapports d'évaluation ou des publications, et au moins une phrase suggère comment résoudre ces divergences. L'approche MM par *tension dialectique* peut coexister avec la précédente. Elle postule l'existence de différentes « visions partielles du monde » (Teddle et Tashakkori, 2003, p. 18). Trois principales stratégies sont proposées pour résoudre les divergences entre données ou résultats QUAL et QUANT : exclusion, réconciliation ou initiation (nouveau projet de recherche ou nouvelle collecte/analyse de données). Ces stratégies sont présentées ailleurs (Pluye, Grad, Levine et Nicolau, 2009).

L'approche MM par *assimilation* suggère que les résultats QUAL et QUANT soient assimilés ou transformés sous une seule forme QUAL ou QUANT. Par exemple, les fréquences d'occurrence des thèmes QUAL sont présentées, et il n'y a aucune distinction claire entre les résultats QUAL et QUANT dans la section « Résultats » des rapports d'évaluation ou des publications.

Les devis

Une classification usuelle des méthodes mixtes est fondée sur trois types de devis de recherche qui correspondent chacun à un type d'intégration des méthodes quantitatives et qualitatives (Creswell et Plano Clark, 2010). Premièrement, les devis de recherche séquentiels exploratoires qui utilisent une méthode qualitative suivie d'une méthode quantitative (QUAL puis QUANT). En termes d'intégration, les résultats quantitatifs sont mobilisés pour confirmer ou généraliser les résultats qualitatifs. Deuxièmement, les devis séquentiels explicatifs qui utilisent une méthode quantitative suivie d'une méthode qualitative (QUANT puis QUAL). Concernant l'intégration, les résultats qualitatifs sont mobilisés pour interpréter ou expliquer des résultats quantitatifs. Troisièmement, les devis concomitants combinent les méthodes quantitatives et qualitatives dès la planification de la recherche, ou au moment de la collecte et de l'analyse des données (QUANT et QUAL). Par exemple, ces méthodes peuvent étudier les mêmes phénomènes de manière complémentaire. L'intégration s'effectue au moment de la collecte et de l'analyse des données.

Ces trois types de devis MM sont proposés pour planifier des études simples plutôt que des programmes de recherche composés de plusieurs

études, mais ils peuvent certainement s'appliquer aux deux (Creswell et Plano Clark, 2010). On peut transposer cette nuance dans le domaine des évaluations de la manière suivante. Les MM en évaluation de programmes peuvent combiner des méthodes QUAL et QUANT à l'intérieur d'une étude d'évaluation unique ou dans un ensemble d'évaluations étroitement liées. À titre d'illustration, nous présentons ici un devis d'évaluation des technologies. Cet exemple montre comment combiner une évaluation QUANT longitudinale (étude de cohorte) avec une évaluation QUAL (étude de cas multiple) dans un devis MM concomitant.

Il existe aussi des variantes de ces trois devis de base, par exemple les devis multiniveaux. Le chapitre de Dagenais et Dupont (chapitre 22) qui présente l'évaluation du programme appelé « Équipe d'intervention jeunesse » en fournit une illustration. Ce programme pilote (2003-2006) comportait l'analyse de la mise en œuvre de 17 équipes de coordination pour améliorer l'accessibilité et la qualité des services pour des jeunes en difficulté (0-18 ans). Trois aspects du programme ont été évalués à deux niveaux, (a) les jeunes (1. implantation et 2. résultats), et (b) les équipes (3. processus de coordination). *Au premier niveau*, l'évaluation a utilisé un devis concomitant (triangulation). La composante QUANT consistait en analyses statistiques descriptives des données collectées pour chaque jeune ($N = 441$), dans une base de données sur internet, par les coordinateurs du programme (étude de prévalence). La composante QUAL consistait en études de cas ($N = 29$), un cas étant défini par un jeune, son entourage et les professionnels chargés de leur fournir des services. Les cas ont été sélectionnés de manière aléatoire de façon à ce qu'ils soient représentatifs de la population des 441 jeunes. Les évaluateurs ont mené 114 entrevues semi-structurées pour collecter des données, et une analyse thématique des transcriptions de ces entrevues. Les résultats QUANT suggèrent que le programme est bien implanté, atteint la clientèle appropriée et fournit comme prévu des services provenant de différentes organisations. Les résultats QUAL suggèrent que les résultats du programme sont positifs en termes de satisfaction des jeunes, de leur famille et des professionnels, par exemple, les parents ont fait part d'une amélioration de l'accès et de la qualité des services offerts. *Au deuxième niveau*, un échantillon de quatre équipes a été sélectionné de manière à refléter les différents types d'équipes en termes de composition, de fonctionnement, et de clientèle. Pour chaque équipe, des données additionnelles ont été

collectées en effectuant des entrevues individuelles avec les coordinateurs, les agents de liaison et les responsables des services, également en menant un *focus group* avec eux. Les évaluateurs ont effectué une analyse thématique des transcriptions des entrevues et des *focus groups*, tout en tenant compte des données de « premier niveau ». Les résultats suggèrent par exemple que le coordinateur avait un rôle de facilitateur central pour le bon fonctionnement des équipes.

Un exemple de devis concomitant

- *Évaluation quantitative* (Grad *et al.*, 2005). Nous avons élaboré une méthode pour évaluer systématiquement l'impact des recherches d'informations dans les ressources électroniques. Nous avons évalué une ressource, InfoRetrieveur, avec une cohorte de 26 résidents. InfoRetrieveur est un engin de recherche qui combine sept bases de données, incluant un manuel de référence, des lignes directrices, des systèmes experts, et des synopsis d'articles de recherche sélectionnés pour leur pertinence et leur validité en médecine familiale. Un questionnaire électronique d'évaluation était rempli par ces professionnels pour rapporter l'impact perçu de chaque item d'information trouvé dans InfoRetrieveur. Le questionnaire a permis d'évaluer 2495 items d'information. Les rapports d'impact positif les plus fréquents étaient « Ma pratique est améliorée » et « J'ai appris quelque chose de nouveau ».
- *Évaluation qualitative concomitante* (Pluye *et al.*, 2007). Notre méthode a permis d'évaluer systématiquement la pertinence situationnelle des ressources électroniques. Nous avons d'abord identifié sept raisons pour lesquelles les professionnels cherchent des informations : quatre objectifs cognitifs (par exemple, répondre à une question clinique) et trois objectifs organisationnels (par exemple, échanger des informations avec d'autres professionnels). Puis, nous avons conduit une étude de cas multiple. Les cas consistaient en des recherches effectuées dans InfoRetrieveur par 17 résidents pendant deux mois. Les entrevues avec ces professionnels (collecte de données qualitatives) étaient guidées par les données du questionnaire mentionné ci-dessus (collecte concomitante de données quantitatives). Notre analyse concomitante des données qualitatives et quantitatives a porté sur 156 recherches critiques (contexte et conséquences clairement décrits) liées à 877 items d'information. Pour chaque cas, les données qualitatives étaient assignées à un objectif cognitif ou organisationnel (analyse thématique). Les résidents ont atteint leurs objectifs dans 85,9 % des cas (pertinence situationnelle). Les résultats soutiennent les objectifs proposés et suggèrent des niveaux de pertinence.
- *Évaluation quantitative modifiée*. Grâce aux résultats du devis concomitant, notre nouveau questionnaire intègre les sept objectifs ci-dessus pour évaluer systématiquement l'impact et la pertinence situationnelle des informations.

Sources : Grad, R. M. *et al.* (2005), « Assessing the impact of clinical information-retrieval technology in a family practice residency », *Journal of Evaluation in Clinical Practice*, vol. 11, n° 6.

Pluye, P. *et al.* (2007), « Seven reasons why family physicians search clinical information-retrieval technology: Toward an organizational model », *Journal of Evaluation in Clinical Practice*, vol. 13, n° 1.

Les techniques

Les techniques mobilisées en MM pour évaluer des programmes peuvent s'inspirer de tous les types de combinaisons des techniques pour la collecte et l'analyse de données QUAL et QUANT. Bryman (2006) a effectué une recension des écrits couvrant 232 articles en sciences sociales qui rapportaient l'emploi de MM. Il en a conclu que l'entrevue structurée, ou questionnaire, tend à prédominer du côté des techniques QUANT (enquêtes transversales), et que l'entrevue ouverte ou semi-structurée tend à prédominer du côté des techniques QUAL. Selon leur synthèse des écrits, les principales techniques de collecte de données associées aux MM sont, par ordre alphabétique : les entrevues individuelles (non structurées, semi-structurées, structurées), les groupes de discussion, les questionnaires (avec ou sans réponses ouvertes), l'observation participante (de type ethnographique) et la recension de documents. Selon cette synthèse des écrits, les principales techniques d'analyse de données associées aux MM sont, par ordre alphabétique : l'analyse thématique qualitative, les analyses statistiques, et l'analyse de contenu ou la quantification de données qualitatives (par exemple, la quantification des réponses fournies par les interviewés).

De plus, une proportion substantielle des articles mentionne que la collecte de données QUANT et QUAL n'était pas fondée, comme on pourrait s'y attendre, sur des instrumentations distinctes. Une telle combinaison des techniques correspond au *Devis concomitant* présenté plus haut. Cette combinaison suppose par exemple que des entrevues ethnographiques peuvent être analysées en utilisant des statistiques textuelles (MM lexico-sémantiques). Cette combinaison de techniques illustre encore une fois comment les MM suggèrent des types méthodologiques différents par rapport aux idéaux-types *empirisme logique* et *constructivisme*. Cet aperçu des approches, devis et techniques permet de planifier les MM en évaluation de programmes, et d'aborder la question de la qualité des MM.

La planification et la qualité des méthodes mixtes en évaluation de programmes

Il existe peu de recommandations spécifiques pour planifier les MM en évaluation de programmes (Greene, Caracelli et Graham, 1989 ; Greene,

2006), mais les évaluateurs peuvent s'inspirer des recommandations existantes pour planifier des recherches MM (Creswell et Plano Clark, 2010). Au moment où nous écrivons ce chapitre, il n'existe toujours pas de norme généralement admise pour planifier une MM de « bonne qualité » ou pour évaluer la qualité méthodologique des MM. Selon Creswell *et al.* (2004) et O'Cathain (2010), des travaux sont nécessaires pour préciser les critères de qualité des MM.

À partir de notre recension des écrits (Pluye, Gagnon, Griffiths et Johnson-Lafleur, 2009), d'une étude pilote (Pace *et al.*, 2012), d'ateliers et de consultations avec des experts (Pluye *et al.*, 2011), nous proposons de concevoir la qualité méthodologique des MM de la manière suivante (tableau 7.2). Les MM doivent réconcilier les idéaux-types *empirisme logique* et *constructivisme* résumés plus haut. La théorie constructionniste de Hacking (1999) permet, à notre avis, cette réconciliation. Les MM créent un « effet de boucle » entre les données probantes provenant des études QUAL et QUANT, et construisent ainsi un nouveau « type mixte » de données probantes. Cette conception des MM soutient l'usage de différents critères pour différentes méthodes. Ainsi, nous proposons : 1) d'évaluer la qualité méthodologique des composantes QUAL des MM à partir de critères utilisés pour évaluer les méthodes QUAL ; 2) d'évaluer les composantes QUANT des MM à partir de critères utilisés en épidémiologie pour évaluer les méthodes QUANT ; et 3) d'évaluer la qualité générale des MM en fonction de ce que nous avons présenté dans la section précédente sur les approches, devis et techniques des MM. Fondé sur cette conception, un outil d'évaluation de la qualité des méthodes mixtes est disponible gratuitement sur internet (<http://mixedmethodsappraisaltool-public.pbworks.com>). Cet outil tient sur une seule page, il est fourni avec un manuel qui en facilite l'utilisation.

Il faut souligner que l'évaluation de la qualité des méthodes QUAL est controversée, et donc, par conséquent, celle des composantes QUAL des MM l'est aussi pour au moins trois raisons : quelle que soit leur qualité, certaines études QUAL peuvent *illuminer*, c'est-à-dire clarifier ou améliorer la compréhension des phénomènes observés (Sandelowski, 1993) ; peu de méthodes QUAL correspondent à des critères de qualité explicites généralement admis par une communauté de chercheurs (Sandelowski *et al.*, 2007) ; enfin, les méthodes QUAL ne sont pas des entités fixes telles que présentées dans les outils d'évaluation de la qualité, et nombreux sont

TABLEAU 7.2**Qualité des méthodes mixtes en évaluation de programmes**

Examiner:	Critères de qualité des méthodes mixtes
1. Composante qualitative	• Sources de données pertinentes pour répondre à la question • Analyse des données pertinentes pour répondre à la question • Prise en compte du contexte de collecte des données • Réflexivité des évaluateurs sur leurs méthode et résultats
2. Composante quantitative randomisée	• Randomisation (ou génération de séquence) appropriée • Allocation appropriée (par ex., à l'aveugle) • Résultats complets • Faible taux d'abandon
3. Composante quantitative non randomisée	• Recrutement minimisant les biais • Mesures appropriées (validées ou standards) • Participants similaires dans les deux groupes (ou différences prises en compte dans l'analyse) • Résultats complets, taux de réponse et de suivi acceptables
4. Composante quantitative descriptive	• Échantillonnage approprié pour répondre à la question • Échantillon représentatif de la population étudiée • Mesures appropriées (validées ou standards) • Taux de réponse acceptable
5. Composante mixte	• Devis MM pertinent pour répondre aux questions • Intégration des données et (ou) résultats QUAL et QUAN • Discussion des limites associées à cette intégration (par ex., divergences QUAL vs QUANT)

ceux qui critiquent la prémisse voulant qu'il existe clairement une bonne et une mauvaise manière d'appliquer les méthodes QUAL.

Par conséquent, alors même que se développent les MM, il faut tenter d'en promouvoir la qualité en s'appuyant sur des critères de rigueur et ne pas tuer dans l'œuf la créativité et les découvertes associées aux composantes qualitatives des MM en imposant des critères stricts qui, justement, empêcheraient de bénéficier de la richesse que peut offrir la combinaison

des méthodes QUAL et QUANT. Pour cela, il serait possible d'utiliser les critères proposés au tableau 7.2 pour analyser séparément, en évaluation de programmes, les études et les composantes QUAL ou QUANT de bonne vs de faible qualité méthodologique.

Pour illustrer l'utilisation de cet outil, nous résumons ici la description et la discussion des méthodes utilisées au chapitre 21 pour évaluer un programme. Nous appliquons ensuite les critères présentés ci-dessus pour en évaluer la qualité méthodologique. Ridde *et al.* (chapitre 21) présentent l'utilisation d'un devis séquentiel explicatif pour (a) mesurer les dépenses effectuées par les femmes lors d'un accouchement eutocique dans une maternité de première ligne d'un district du Burkina Faso au moyen d'une enquête de prévalence, puis (b) comprendre ces résultats au moyen d'une étude de cas qualitative. Ils ont présenté les résultats quantitatifs sur l'écart constaté entre le forfait officiel (norme) et les paiements déclarés par les femmes (pratique), puis les résultats qualitatifs pour comprendre cet écart. En plus des leçons apprises de ces résultats pour améliorer la politique nationale en soins de santé primaire, les auteurs suggèrent des recommandations pour limiter l'écart de mise en œuvre de politiques publiques similaires (censées favoriser l'équité).

Composante quantitative descriptive: En conduisant une enquête auprès des ménages, l'échantillonnage était approprié pour répondre aux questions sur les dépenses. L'échantillon était représentatif de la population étudiée puisqu'il incluait quasiment toute la population (98,7 % des femmes ayant accouché et enregistrées dans les maternités). Les auteurs mentionnent que le questionnaire était adapté d'une précédente recherche. Pour être systématique et vérifier que les mesures étaient appropriées, nous pourrions contacter les auteurs et demander des précisions sur la validité et l'adaptation du questionnaire.

Composante qualitative: Les sources de données étaient pertinentes pour répondre à la question (*focus groups* et entrevues individuelles). Les auteurs justifient de manière convaincante les raisons du choix des informateurs clés parmi les agents de première ligne dans les centres de santé, les responsables des maternités ou des centres de santé et les membres de l'équipe cadre du district. Les résultats quantitatifs étaient présentés par les chercheurs et une discussion était menée pour examiner les pistes d'explications. L'analyse thématique inductive des données était pertinente pour répondre à la question. Les auteurs expliquent comment ils ont pris

en compte le contexte de la collecte des données, et ils fournissent une description détaillée du site, de la population, de l'organisation pyramidale (hiérarchique) de la société et du système de santé (pouvoir important des professionnels), et de la mise en œuvre du programme évalué. La réflexivité des évaluateurs sur leur méthode et leurs résultats est largement discutée. Par exemple, comment les biais inhérents aux discussions de groupe (contrôle social) ont été diminués en ayant également recours à des entrevues individuelles en profondeur qui ont été conduites dans des maternités identifiées comme « cas extrêmes » grâce à l'analyse des données quantitatives (ayant la moyenne la plus élevée en termes de paiement déclaré par les femmes).

Composante mixte : Le devis mixte était pertinent pour répondre aux questions. Le devis séquentiel explicatif est justifié de la manière suivante : après l'analyse des données quantitatives sur les dépenses effectuées par les femmes, quatre pistes qualitatives d'explications à l'écart constaté entre le forfait officiel et le paiement déclaré par les femmes ont été formulées. L'intégration des données s'est faite entre les étapes quantitative et qualitative (sélection de cas extrêmes), puis les résultats qualitatifs ont été intégrés aux résultats quantitatifs au moment de l'interprétation des résultats par les chercheurs. Finalement, les auteurs ont discuté les limites associées à cette intégration en termes de forces et faiblesses respectives des composantes quantitatives et qualitatives, par exemple sur l'absence de mesure de l'importance et de la signification du lien de causalité entre les écarts observés et les explications de ces écarts.

Une recension mixte des écrits pour les évaluations de programme

Dans toute démarche d'évaluation, la recension des écrits constitue un élément essentiel qui permet de préciser les objectifs de l'évaluation en fonction de ce qui est déjà connu, ou de discuter de ses résultats. On connaît les métasynthèses d'études QUAL (par exemple, les méta-ethnographies) et les revues systématiques d'études QUANT (par exemple, les revues des essais contrôlés randomisés avec méta-analyse). On connaît cependant moins ce que nous appelons les *recensions mixtes des écrits*, c'est-à-dire les revues concomitantes d'études QUAL, QUANT et MM (Pluye, Gagnon *et al.*, 2009). Par exemple, effectuer une recension mixte permet de faire la synthèse de tout type de connaissances empiriques entourant les

programmes à évaluer ou en cours d'évaluation. Des exemples peuvent être trouvés en éducation, en santé, en gestion et en sciences sociales. Nous résumons en encadré une recension mixte que nous avons effectuée pour améliorer notre méthode d'évaluation des technologies.

Nous avons également examiné un échantillon de 59 recensions mixtes des écrits en sciences de la santé. Ces recensions constituent un type de MM dont les données sont documentaires (documents rapportant des évaluations ou des études tant QUANT, QUAL que MM). Chaque recension mixte examine en moyenne 26 études QUANT, 17,9 études QUAL et 3,7 études MM. Les trois approches mentionnées plus haut sont utilisées par les recensions mixtes. Les résultats QUAL et QUANT sont complémentaires dans 54 % des cas, ils sont en tension dialectique dans 7 % des cas, et sont assimilés à ou transformés en résultats QUANT ou QUAL dans 36 % des cas.

Bénéficiant du développement des MM, les recensions mixtes des écrits ont récemment émergé comme une nouvelle forme de recension des écrits. Des 59 recensions examinées dans le domaine de la santé, 56 (95 %) ont été publiées depuis 2000. Selon notre échantillon, les recensions mixtes sont utiles pour explorer de nouvelles connaissances (par exemple, découvrir les tendances actuelles concernant le programme évalué) et (ou) pour confirmer des savoirs (par exemple, justifier les prémisses du programme évalué). Les recensions mixtes sont utilisées dans des buts exploratoires (62 %) ou de validation (19 %), ou encore dans ces deux buts à la fois (19 %).

Nous avons identifié trois types de recensions mixtes (*systématiques*, *reproductibles* ou *de convenance*) selon qu'elles satisfont ou non aux caractéristiques des recensions systématiques des écrits (Moynihan, 2002) : formulation d'une question ; identification des études potentiellement pertinentes ; sélection des études pertinentes ; évaluation de la qualité des études sélectionnées ; synthèse des résultats des études retenues pour leur pertinence et leur qualité. Nous avons identifié 17 *recensions systématiques* (29 %) qui satisfont aux cinq caractéristiques mentionnées, et 23 (39 %) *recensions reproductibles* qui satisfont quatre caractéristiques (question, identification, sélection et synthèse) et n'évaluent pas la qualité des études sélectionnées pour leur pertinence. Quant aux 19 *recensions de convenance* (32 %), elles satisfont uniquement à deux caractéristiques (question et synthèse), car l'identification et la sélection ne peuvent pas être reproduites

à partir des informations disponibles et parce qu'il n'y a pas d'évaluation de la qualité des études retenues. Si les évaluateurs décident de conduire une recension systématique, ils peuvent utiliser le tableau 7.2 pour évaluer la qualité des études QUAL, QUANT et MM sélectionnées pour leur pertinence.

Une recension mixte des écrits en évaluation des technologies

- L'évaluation des technologies est complexe et ouverte au pluralisme méthodologique. Nous avons identifié, sélectionné, évalué et fait la synthèse de tous les types d'études empiriques examinant l'impact des ressources électroniques sur les professionnels.
- Méthodes. Nous avons recensé les écrits scientifiques internationaux des années 1960 à 2004. Deux réviseurs ont identifié indépendamment les études potentiellement pertinentes en scrutant respectivement 3 368 et 3 249 références (titre, auteurs, revue, résumé) en utilisant deux stratégies de recherche dans des bases de données bibliographiques. Des études additionnelles ont été identifiées à la main (recherche dans des dossiers personnels et des revues spécialisées) et en effectuant une recherche des articles citant les articles sélectionnés par l'entremise de la base de données bibliographique ISI Web of Sciences. Les textes entiers de 605 articles ont été examinés pour leur pertinence. Parmi ceux-ci, 40 (6,6 %) ont été évalués indépendamment par deux réviseurs pour leur pertinence et leur qualité méthodologique. Ces articles rapportaient des recherches évaluatives quantitatives, qualitatives ou mixtes. De ces derniers, 26 (4,3 %) ont été retenus comme pertinents et de qualité acceptable pour faire une synthèse. Pour chaque article retenu, deux équipes ont effectué une analyse thématique qualitative pour faire la synthèse des résultats quantitatifs et qualitatifs.
- Résultats. Les études quantitatives observationnelles suggèrent qu'environ un tiers des recherches d'informations dans les ressources électroniques ont un impact positif sur les médecins. Les études quantitatives expérimentales ou effectuées en laboratoires informatiques (simulations) sont contradictoires quant à l'impact des ressources électroniques par rapport aux autres sources d'information comme le matériel éducatif imprimé. Les résultats qualitatifs et la qualification des résultats quantitatifs soutiennent et complètent les types d'impact que nous avons proposés dans une étude pilote.

Source : Pluye, P. *et al.* (2005), « The Impact of Clinical Information-Retrieval Technology on Physicians: A Literature Review of Quantitative, Qualitative and Mixed-method Studies », *International Journal of Medical Informatics*, vol. 74, n° 9.

Conclusion

Les chercheurs et les évaluateurs qui utilisent les méthodes mixtes rejettent la thèse d'incommensurabilité qui sépare le constructivisme et l'empirisme logique en deux idéaux-types méthodologiques incompatibles (Johnson, 2007). Le présent chapitre montre comment les propositions de Hacking (constructionnisme) peuvent aider à conceptualiser les MM et les données probantes mixtes, et à partir de cette conceptualisation, il décrit trois approches méthodologiques MM pour l'évaluation des programmes, trois types de devis MM et diverses combinaisons de techniques de collecte et d'analyse de données pour les MM. De plus, il propose des

critères pour évaluer la qualité des MM. Selon notre expérience, les MM devraient être planifiées, implantées et évaluées par un expert en méthodes QUAL et QUANT, ou par au moins deux experts ayant des expertises complémentaires en méthodes QUAL et QUANT. Par exemple, notre évaluation des techniques et notre recension des écrits ont été conduites par un chercheur ayant une formation en épidémiologie avec une expertise en MM et en méthodes QUAL, et par des chercheurs détenant une expertise en méthodes QUANT.

En conclusion, nous proposons un plan de cours pour enseigner les MM qui peut être utile aux évaluateurs, aux étudiants et aux professeurs dans le domaine de l'évaluation de programmes. Ce cours a été implanté en 2008. Bien que celui-ci soit destiné aux étudiants dans le domaine de la santé, il a été construit sur la base d'autres programmes nord-américains en éducation, en psychologie et en sociologie. Le cours est conduit par un chercheur QUAL (sociologue), un chercheur QUANT (épidémiologiste) et un chercheur détenant une formation en épidémiologie avec une expertise en MM et en méthodes QUAL.

Enseigner les méthodes mixtes : plan de cours

- Cours 1. Introduction : Afin de préparer les étudiants pour les études mixtes, une brève introduction à chaque type de recherche est fournie aux étudiants.
- Cours 2. Projet de recherche : Les étudiants ébauchent un projet de recherche utilisant les méthodes mixtes.
- Cours 3. Recension mixte des écrits : Les étudiants apprennent les techniques pour conduire une analyse documentaire incluant des études qualitatives, quantitatives et mixtes.
- Cours 4 et 5. Devis mixtes : Les devis mixtes sont présentés, et chaque étudiant applique un devis approprié pour son projet.
- Cours 6. Échantillonnage et instrumentation : Les étudiants construisent l'échantillonnage et l'instrumentation appropriés, avec des composantes qualitatives et quantitatives, pour appliquer leur devis de recherche.
- Cours 7. Collecte de données et réflexion sur l'instrumentation : Les étudiants appliquent et présentent leur instrumentation à leurs pairs et en discutent avec eux.
- Cours 8. Divergences : Les étudiants utilisent trois stratégies pour analyser les divergences entre données qualitatives et quantitatives.
- Cours 9. Analyser des mots et des nombres : Les étudiants organisent et analysent leurs données qualitatives et quantitatives.
- Cours 10. Évaluation de la qualité méthodologique des études utilisant les méthodes mixtes : Les différentes méthodes pour évaluer la qualité des études qualitatives, quantitatives et mixtes sont présentées.
- Cours 11. Session de lecture critique : Les étudiants évaluent de façon critique un article qui utilise un devis mixte.
- Cours 12 : Projets individuels : Les étudiants présentent leurs projets (communication orale) et reçoivent les commentaires et suggestions des instructeurs et de leurs pairs.

Source : Levine, A., B. Nicolau et P. Pluye, « Applied Mixed Methods in Health Research », plan de cours, Université McGill.

Pour aller plus loin :

CRESWELL, J. W. et V. L. PLANO CLARK (2010), *Designing and conducting Mixed Methods Research*, Thousand Oaks, Sage Publications.

Ce livre contient les plus récentes propositions de John Creswell (auteur du livre à succès *Research Design*) et de Vicki Plano Clark sur les types de devis de recherche et d'évaluation utilisant des méthodes mixtes. Il comprend des recommandations pratiques pour élaborer, étape par étape, un devis d'évaluation utilisant des méthodes mixtes. Ce livre constitue un manuel de base suffisant pour mener à bien une évaluation mixte.

GREENE, J. C., V. J. CARACELLI et W. F. GRAHAM (1989), « Toward a conceptual framework for mixed-method evaluation designs », *Educational Evaluation and Policy Analysis*, vol. 11, n° 3, p. 255-274.

Dans cet article pionnier, un cadre conceptuel des méthodes mixtes en évaluation est proposé à partir de la littérature et de l'analyse de 57 recherches évaluatives utilisant des méthodes mixtes. Cinq stratégies d'évaluation utilisant des méthodes mixtes sont identifiées : la triangulation, la complémentarité, le développement, l'initiation et l'expansion.

TASHAKKORI, A. et C. TEDDLIE (2010), *Handbook of Mixed Methods in Social and Behavioral Research*, Thousand Oaks, Sage Publications.

Ce livre présente les fondements de l'utilisation des méthodes mixtes. Les défis théoriques sont très bien développés et illustrés à l'aide d'exemples concrets. Tous les aspects des méthodes mixtes sont abordés par les auteurs clés dans le domaine. L'ouvrage offre une synthèse des méthodes mixtes très complète. Par rapport au manuel de Creswell et Plano Clark (2010), il constitue une référence incontournable pour les évaluateurs, les étudiants et les chercheurs qui veulent utiliser les méthodes mixtes de manière plus avancée.

Internet

Mixed Methods for Health Research: Overview of mixed methods research design & review.

Cet ouvrage, libre d'accès en ligne, contient une liste de ressources mises à jour par Frances Griffiths et Felicity Boardman de l'Université de Warwick (glossaire, bibliographie, experts, cours et formation, logiciels, sites internet, conférences et réseaux) :

www2.warwick.ac.uk/fac/med/research/hsri/primary_care/research_/centrepatexp/complexityhealth/mixedmethods

Journal of Mixed Methods Research.

La publication internationale de référence pour les articles qui portent sur les méthodes mixtes dans toutes les disciplines. Soutenu par les chefs de file de ce domaine, chaque numéro explore des recherches et des évaluations originales utilisant des méthodes mixtes, et présente des articles méthodologiques et théoriques qui contribuent aux connaissances sur ces méthodes. Le site de la revue : www.sagepub.com/journalsProdDesc.nav?prodId=Journal 201775

8

L'évaluation axée sur l'utilisation

Michael Quinn Patton et Francine LaBossière

L'évaluation axée sur l'utilisation est basée sur le principe selon lequel les évaluations devraient être jugées en fonction de leur utilité. Ainsi, un évaluateur devrait concevoir et mener une évaluation en tenant compte, du début jusqu'à la fin, de son utilisation future. Pour ce faire, celui-ci doit bien comprendre ce que les demandeurs d'une évaluation et les utilisateurs ciblés veulent savoir et comment ils entendent disposer des informations qui en découleront. L'accent est donc mis sur l'*utilisation attendue* des résultats de l'évaluation par des personnes identifiées à l'avance, les *utilisateurs attendus*.

Pour toute évaluation, il y a plusieurs utilisateurs potentiels et diverses utilisations possibles. L'évaluation axée sur l'utilisation exige que l'on passe de ce qui est général et vague – c'est-à-dire de publics potentiels et d'utilisations possibles – à ce qui est spécifique et concret : des usagers bien identifiés et un engagement manifeste de leur part à se servir de l'évaluation à des fins précises. L'utilité d'une évaluation est trop importante pour n'être qu'espérée ou présumée ; il est donc nécessaire de la prévoir et de la renforcer tout au long du processus évaluatif.

Une erreur souvent commise est de présumer qu'une évaluation contribuera forcément à l'amélioration d'un programme ou aux décisions qui l'entourent. Pourtant, les décideurs et les intervenants du programme ne sont pas nécessairement enchantés lorsqu'ils reçoivent les résultats d'une évaluation. Parmi les critiques les plus fréquentes, on retrouve celles-ci :

- le rapport est en retard, les décisions ont déjà été prises ;
- le rapport est très volumineux, personne ne le lira ;
- les questions posées ne sont pas les bonnes ;
- le rapport ne nous dit pas ce que nous voulions savoir ;
- le jargon rend le rapport ennuyeux et difficile à comprendre.

L'évaluation axée sur l'utilisation ne prône pas un seul type d'évaluation, de modèle, de méthode, de théorie ni même d'utilisation. Il s'agit plutôt d'un processus visant à aider les utilisateurs de l'évaluation à faire ces choix en fonction de leurs besoins. Comme on peut le voir dans les différents chapitres de ce manuel, il existe aujourd'hui plusieurs options dans le domaine de l'évaluation. L'approche axée sur l'utilisation peut s'appliquer à tout type de but, d'objet d'évaluation, de devis et de données. Quel que soit le type d'évaluation choisi, cette approche vise à en assurer une utilisation optimale.

Comment est-il possible d'accroître l'utilité d'une évaluation ? En posant les bonnes questions, en sachant quels sont les vrais besoins d'information, en faisant coïncider les résultats de l'évaluation avec les moments de prise de décision, et en travaillant étroitement avec les utilisateurs attendus de l'évaluation.

La clé d'une évaluation utile est de bien préciser les besoins d'information et de s'interroger sur la pertinence de l'évaluation tout au long du processus : *Pourquoi fait-on cela ? À quoi cela servira-t-il ? Est-ce vraiment ce que l'on veut savoir ? Cette façon de faire nous donnera-t-elle des informations utiles ? Comment compte-t-on concrètement utiliser les données ?* Si les utilisateurs attendus prévoient se servir des résultats pour « améliorer une intervention », il faut les aider à être plus précis : *Veut-on un programme plus efficace ? Mieux perçu par la population ? Qui dessert une plus grande clientèle ? Quels aspects de l'intervention veut-on améliorer ?* En posant ce genre de questions, l'évaluateur et les utilisateurs augmentent les chances de voir l'évaluation répondre efficacement à leurs besoins et être réellement utile.

Si une évaluation a été demandée en vue d'aider à une prise de décision, il faut savoir quelles sont les questions que les décideurs se posent vraiment.

Veulent-ils savoir si un programme local devrait être étendu à un plus grand nombre de régions ? Est-ce qu'ils veulent connaître les effets d'un

programme sur quelques indicateurs spécifiques? Ou alors, dans un contexte de restrictions budgétaires, la véritable question porte-t-elle sur la possibilité de faire aussi bien avec moins? En connaissant les vraies questions et le contexte dans lequel une décision doit être prise, un évaluateur peut s'assurer d'aller chercher des données qui seront pertinentes pour les décideurs.

Cependant, il ne suffit pas de donner les *bonnes informations* aux décideurs; il faut aussi les leur transmettre au *bon moment*. Un évaluateur doit donc savoir quand se prendront les décisions et s'assurer que les informations voulues seront disponibles au moment opportun. Sans ce souci constant de la pertinence et des échéances, les décideurs risquent de se retrouver face à des résultats d'évaluation qu'ils pourront juger intéressants, mais qui ne les aideront pas dans leur prise de décision.

Enfin, la recherche en psychologie a montré qu'une personne sera plus apte à se servir d'une évaluation si elle y a participé activement (Patton, 1997). Une participation active favorise l'appropriation de l'évaluation et une meilleure compréhension des résultats, ce qui augmente la probabilité que ceux-ci soient utilisés. En travaillant étroitement avec les utilisateurs tout au long du processus, l'évaluateur prépare le terrain pour un usage concret des résultats et renforce l'utilité de l'évaluation.

Le facteur personnel

Au cours d'une évaluation, plusieurs choix importants doivent être faits. Entre autres, il faut déterminer le but de l'évaluation, choisir les méthodes à appliquer et s'entendre sur les échéanciers. Qui sont les personnes qui doivent faire ces choix? Dans l'évaluation axée sur l'utilisation, ce sont les *principaux utilisateurs attendus de l'évaluation* qui doivent prendre ces décisions, avec le soutien de l'évaluateur.

Tout individu qui est enclin à utiliser une évaluation et qui est en mesure de le faire peut être considéré utilisateur d'une évaluation. Les *principaux utilisateurs attendus* sont les personnes ou les groupes qui sont particulièrement concernés par une évaluation, qui ont le pouvoir et la capacité d'y contribuer activement, et qui ont l'intention de l'utiliser à des fins concrètes. Ils collaborent avec l'évaluateur tout au long du processus pour s'assurer que l'évaluation est pertinente et qu'elle répond à leurs besoins. Ces personnes sont considérées comme *parties prenantes de l'évaluation*, bien qu'elles n'en soient pas les seules.

Les parties prenantes d'une évaluation sont les personnes et les organismes pour qui les résultats de l'évaluation présentent un intérêt particulier. Toute personne qui doit prendre des décisions concernant le programme ou qui désire obtenir de l'information sur celui-ci peut être considérée partie prenante. Il s'agit entre autres des bailleurs de fonds, des gestionnaires du programme, du personnel et des bénéficiaires. D'autres personnes ayant un intérêt indirect dans l'efficacité du programme peuvent aussi être partie prenante de l'évaluation; cela peut inclure les médias, les chercheurs ou encore les contribuables, dans le cas d'un programme public. Cependant, les différentes parties prenantes peuvent avoir des intérêts très variés et parfois contradictoires. Puisque aucune évaluation ne peut répondre adéquatement à tous leurs besoins d'information, il faut donc adopter un processus permettant de choisir les questions d'évaluation prioritaires. Dans l'évaluation axée sur l'utilisation, ce processus commence en repérant les utilisateurs potentiels et en identifiant, parmi ceux-ci, les principaux utilisateurs attendus. Leurs besoins d'information spécifiques orienteront l'évaluation.

L'évaluation axée sur l'utilisation débute donc par l'identification explicite d'une personne ou d'un groupe de personnes spécifiques qui s'engagent à utiliser une évaluation à des fins précises. Ceci diffère de ce qui a traditionnellement été recommandé, c'est-à-dire d'identifier « les publics » visés par une évaluation. Un public est une entité plutôt vague et anonyme, et dont on ne peut être certain de l'engagement envers l'usage de l'évaluation. Il ne suffit donc pas d'identifier une agence ou une organisation en tant que destinataire du rapport d'évaluation. Ce sont les personnes, et non les organisations, qui utilisent les résultats des évaluations, d'où l'importance du *facteur personnel*.

Le « facteur personnel » désigne l'existence d'un individu ou d'un groupe de personnes spécifiques qui s'intéressent personnellement à l'évaluation et à ses résultats. Ces acteurs se sentent interpellés par l'évaluation et y jouent un rôle actif. L'évaluateur axé sur l'utilisation facilite cet engagement, qui se traduit généralement par des utilisations concrètes. Lorsque le facteur personnel est manquant, l'impact de l'évaluation peut être beaucoup moindre.

Les étapes d'une évaluation axée sur l'utilisation

Les cinq étapes principales d'une évaluation axée sur l'utilisation sont les suivantes : 1) identifier les principaux utilisateurs attendus ; 2) préciser les objectifs et les utilisations attendues de l'évaluation ; 3) choisir les méthodes ; 4) analyser et interpréter les données ; et 5) utiliser et diffuser les résultats.

1. Premièrement, il faut donc déterminer les principaux utilisateurs attendus d'une évaluation. Bien sûr, les demandeurs d'une évaluation figurent parmi ceux-ci, mais est-ce qu'il y en a d'autres ? Une analyse soignée des parties prenantes de l'évaluation et des utilisateurs potentiels devrait servir de base à cet exercice. Une fois que les utilisateurs attendus ont été identifiés, il faut déterminer leur niveau de participation approprié à l'évaluation (voir tableau 8.1). Ces décisions se prennent normalement par les demandeurs de l'évaluation – guidés par l'évaluateur – en tenant compte des considérations politiques, éthiques et pratiques propres à chaque situation. Finalement, il faut décider comment les principaux utilisateurs attendus travailleront avec l'évaluateur et participeront à la prise de décision sur l'évaluation (par exemple, dans un groupe de travail).
2. Deuxièmement, l'évaluateur et les principaux utilisateurs attendus s'entendent sur le but de l'évaluation et sur ses utilisations prioritaires. L'évaluateur travaille alors avec les utilisateurs pour déterminer leurs besoins d'informations spécifiques et pour préciser les objectifs et les questions d'évaluation.
3. La troisième étape du processus porte sur les choix de devis, de méthodes et d'outils. Les principaux utilisateurs attendus participent à la prise de décision afin de bien comprendre les forces et les faiblesses des résultats d'évaluation qu'ils auront éventuellement à utiliser. Plusieurs options méthodologiques peuvent être considérées et l'évaluateur fait ressortir les avantages et désavantages de chacun. La discussion à cette étape porte sur les questions de justesse méthodologique, de validité des données, de faisabilité, de bien-fondé et de coûts. Comme toujours, le souci qui prime est celui de l'utilité. Est-ce que les résultats obtenus par ces méthodes seront utiles – et réellement utilisés ?
4. Une fois que les données ont été collectées et compilées, la quatrième étape du processus axé sur l'utilisation commence. Les utilisateurs participent activement à l'interprétation des données et à l'élaboration de

TABLEAU 8.1
Les degrés de participation des différents utilisateurs attendus

Degré de participation					
	Information	Consultation	Implication	Collaboration	Empowerment
LA PROMESSE DE L'ÉVALUATEUR	Nous vous tiendrons au courant du progrès de l'évaluation et des résultats obtenus.	Nous vous tiendrons au courant du déroulement de l'évaluation, nous vous écouterons et nous vous dirons comment vos commentaires et suggestions ont influencé l'évaluation.	Nous travaillerons avec vous pour nous assurer que vos préoccupations soient prises en compte et reflétées dans les options considérées. Nous nous assurerons que vous pourrez revoir et commenter les options possibles et nous vous dirons comment vos commentaires ont été pris en compte dans l'évaluation.	Nous incorporons au maximum vos conseils et suggestions dans l'évaluation. Nous vous donnerons des occasions significatives pour prendre part au processus de prise de décision.	Cette évaluation est la vôtre. Nous vous offrons des options pour que vous puissiez nous transmettre vos décisions. Vous prendrez les décisions et nous vous appuierons dans leur mise en œuvre.
RÔLES ATTENDUS DES UTILISATEURS	Diffuser les résultats de l'évaluation et susciter de l'intérêt pour ceux-ci.	Anticiper les enjeux possibles, identifier les obstacles, suggérer des priorités et rehausser la crédibilité de l'évaluation.	Affirmer l'importance, la justesse et l'utilité de l'évaluation; attirer l'attention sur les résultats et en établir la crédibilité.	Agir en tant que <i>principaux</i> utilisateurs attendus, étant donné leur intérêt élevé dans l'évaluation, leur disponibilité, leurs positions ou réseaux d'influence et leur sentiment d'appropriation par rapport à l'évaluation.	Mener et s'approprier l'évaluation, en plus de développer leurs capacités. L'évaluation est utilisée pour renforcer leur capacité à s'engager dans la pensée et la pratique évaluative.

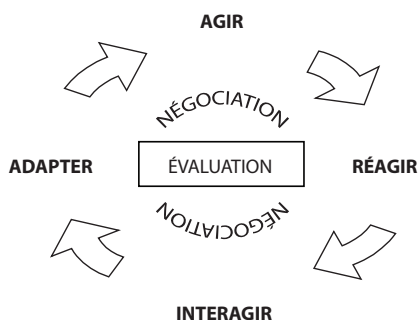
recommandations. Avec l'évaluateur, ils peuvent formuler des stratégies spécifiques pour optimiser l'utilisation de l'évaluation compte tenu des résultats réels qu'ils ont obtenus.

5. Finalement, les utilisateurs attendus de l'évaluation utilisent les résultats aux fins prévues. Au-delà de l'engagement initial et des utilisations qui ont été planifiées, une diffusion plus générale des résultats peut engendrer des utilisations espérées mais non nécessairement attendues.

En théorie, l'évaluation axée sur l'utilisation est un processus relativement simple qui se fait étape par étape. Dans la réalité, il est rarement aussi simple et linéaire. Par exemple, l'évaluateur peut constater que de nouveaux utilisateurs deviennent importants ou que de nouvelles questions d'évaluation émergent lors de la prise de décision sur les méthodes d'évaluation. D'autre part, il n'y a pas toujours de distinction précise entre le choix des questions d'évaluation et le choix des méthodes; les questions guident le choix des méthodes, mais les préférences méthodologiques ou la disponibilité des données peuvent à leur tour guider le choix des questions.

La plupart des décideurs se préoccupent beaucoup de la qualité d'une évaluation; ils veulent qu'elle soit « bien faite ». Ils vont parfois hésiter à contribuer activement à sa réalisation parce que, justement, ils disent ne pas avoir les connaissances pour « bien le faire ». La croyance selon laquelle il n'existe qu'une seule bonne façon de faire les choses est parfois difficile à défaire. Dans une évaluation axée sur l'utilisation, la bonne façon est celle qui sera utile et pertinente pour les utilisateurs attendus. Pour y parvenir, il faut de la négociation et de l'analyse situationnelle.

L'expression « actif-réactif-interactif-adaptatif » décrit l'approche qu'un évaluateur axé sur l'utilisation adopte pour s'assurer qu'une évaluation est utile. Il est en premier lieu *actif* lorsqu'il travaille avec les demandeurs de l'évaluation pour préciser le but de l'évaluation et pour cibler les principaux utilisateurs attendus. Il est *réactif* quand il écoute les utilisateurs et qu'il réagit à ce qu'il apprend sur le contexte particulier dans lequel l'évaluation se déroule. Il est *interactif* lorsqu'il favorise un dialogue et un processus de négociation ouvert. Enfin, il est *adaptatif* quand il permet la modification du plan d'évaluation en fonction de sa compréhension grandissante de la situation et des conditions qui évoluent. Un évaluateur actif-réactif-interactif-adaptatif n'impose pas de recettes d'évaluation préétablies et il ne fait pas la même chose dans chacune de ses évaluations.

FIGURE 8.1**Travailler avec les principaux utilisateurs attendus : un cycle dynamique**

Il est véritablement immergé dans les défis que présente chaque nouvelle situation et il répond de façon personnalisée aux utilisateurs de chaque nouvelle évaluation.

La démarche actif-réactif-interactif-adaptatif reconnaît de façon explicite l'importance de l'expérience, des connaissances et de la contribution de l'évaluateur en mettant l'accent sur l'*action*. Un évaluateur doit être sensible à la situation, mais cela ne veut pas dire qu'il se laisse manipuler par les parties prenantes d'une évaluation, leurs intérêts et leurs besoins. Tout comme un évaluateur axé sur l'utilisation ne doit pas imposer unilatéralement un but et un ensemble de méthodes à une évaluation, les utilisateurs ne doivent pas imposer leurs préférences de façon unilatérale. Pour en arriver à un plan d'évaluation final, il faut un processus de négociation qui permet d'intégrer les valeurs et les capacités de l'évaluateur à celles des utilisateurs attendus.

L'utilité du processus d'évaluation

Généralement, les discussions sur l'utilité des évaluations portent seulement sur l'utilité des résultats. Cependant, le simple acte de participer au processus peut aussi être utile. Si le fait d'avoir pris part activement à l'exercice amène les participants à raisonner comme un évaluateur et à adhérer aux valeurs propres à la discipline, l'évaluation a donc généré plus que des résultats.

Ceux qui ont une formation en méthodes de recherche ou en évaluation peuvent facilement tenir pour acquise la logique qui les sous-tend. En

revanche, pour des décideurs et des gens du terrain, cette logique peut être plus difficile à comprendre ou à appliquer. Penser en fonction de ce qui est spécifique et mesurable n'est pas nécessairement facile pour ceux qui doivent souvent se contenter d'imprécisions et de données informelles pour passer à l'action. Par exemple, lorsqu'un programme tire à sa fin et qu'il faut déterminer son avenir, les décideurs ont à se prononcer même s'ils n'ont pas à leur disposition toutes les informations qu'ils auraient voulues ou des données probantes sur les effets du programme. Les données d'évaluation ne sont qu'un élément parmi plusieurs qu'ils prennent en considération pour prendre une décision, et leur capacité à utiliser ces données dans la prise de décision peut être très variable.

Le processus d'évaluation offre de nombreuses occasions d'apprentissage à ceux qui y participent. Ceux-ci peuvent développer leur capacité à trouver des données probantes, à évaluer des conclusions d'études et à les utiliser plus aisément dans la prise de décision. Ces compétences vont bien au-delà de l'évaluation en cours et pourront leur être utiles tout au long de leur vie professionnelle. L'exercice peut aussi les amener à développer une pensée plus analytique, à se poser de nouvelles questions sur leurs pratiques d'intervention, et à promouvoir une gestion éclairée par les données probantes au sein de leurs organismes respectifs. Bref, le fait de se familiariser avec la démarche évaluative peut avoir un impact significatif sur les participants, un impact qui peut durer plus longtemps que les résultats qui découlent de l'évaluation.

L'utilité du processus d'évaluation ne se limite toutefois pas au renforcement des capacités et au développement d'une pensée évaluative chez les gens qui y participent. Un processus d'évaluation participatif peut aussi, entre autres, favoriser une compréhension partagée d'un programme entre les différents partenaires; augmenter l'engagement et le sens d'appropriation des participants; encourager un dialogue ouvert entre les décideurs et les gens du terrain; rehausser la crédibilité d'un programme; et contribuer au développement des organisations participant à l'évaluation.

Bien sûr, toute évaluation peut avoir ce genre d'effet et c'est d'ailleurs souvent le cas. Mais dans l'évaluation axée sur l'utilisation, la participation active des utilisateurs augmente la probabilité que ces impacts se produisent. On ne les considère pas que comme des retombées informelles; au contraire, on porte une attention particulière à l'utilité du processus et on peut même en faire un but explicite de l'évaluation.

Les principes de l'évaluation axée sur l'utilisation

Un certain nombre de principes sous-tendent l'approche de l'évaluation axée sur l'utilisation. Nous pouvons les résumer de la façon suivante.

- La force motrice d'une évaluation, celle qui guide tout le processus, doit être l'engagement ferme des utilisateurs attendus à s'en servir. Au moment de prendre chacune des décisions – qu'elles concernent les questions, le devis, les méthodes, l'analyse ou le rapport –, l'évaluateur examine avec les utilisateurs comment la décision affectera leur usage de l'évaluation.
- Le souci de l'utilisation est permanent et continu, et ce, dès le début de l'évaluation. Elle n'est pas une chose à laquelle on ne s'intéresse qu'à la fin du processus, le potentiel d'utilisation ayant déjà été largement déterminé une fois que l'évaluation tire à sa fin.
- Les évaluations doivent être spécifiquement orientées vers les intérêts et les besoins d'information de personnes précises, et non de clientèles vagues.
- Les contraintes de temps et de ressources font généralement en sorte qu'une évaluation ne peut pas répondre aux questions de tous ceux qui ont un intérêt dans le programme. Les principales parties prenantes doivent donc négocier le choix des questions prioritaires.
- Toute information n'est pas nécessairement utile. Pour être utile, l'information doit être crédible, pertinente et présentée sous une forme accessible et compréhensible aux utilisateurs.
- Quand l'information est bien ciblée, il y a de meilleures chances qu'elle soit utilisée. Il peut être difficile de savoir précisément à l'avance quelle information sera la plus utile pour la prise de décision ; l'évaluation axée sur l'utilisation vise donc à augmenter la probabilité que l'information recueillie soit pertinente en se focalisant sur les décisions à prendre et en tenant compte des échéances réelles. De cette façon, l'évaluation axée sur l'utilisation minimise l'écart entre l'utilisation potentielle et l'utilisation réelle, et entre les connaissances et l'action.
- Les gens diffèrent dans leur capacité et leur volonté à utiliser l'information. Cela signifie que l'information est plus puissante dans les mains de ceux qui savent l'utiliser et qui sont favorables à son utilisation. Le défi est donc de faire parvenir les bonnes informations aux bonnes personnes.

- Les utilisateurs attendus d'une évaluation sont plus aptes à réellement s'en servir s'ils y ont contribué activement et s'ils en comprennent les différentes étapes.
- C'est la qualité de la participation, et non son volume, qui est importante. Le temps passé en réunion avec les utilisateurs peut être inversement proportionnel à la qualité du processus, si ces rencontres sont inefficaces. Les évaluateurs qui entreprennent des évaluations axées sur l'utilisation doivent être habiles dans l'animation de groupe.
- Une évaluation doit être personnalisée et adaptée au contexte. Les approches basées sur des modèles préétablis ne fonctionnent pas.
- Les éléments qui peuvent nuire à l'utilité des résultats d'évaluation sont tout aussi importants que ceux qui peuvent nuire à leur validité. Les décideurs veulent des données qui sont à la fois robustes *et* utiles.
- Un évaluateur ne doit pas présumer que l'individu ou l'organisme ayant demandé l'évaluation s'engage automatiquement à l'utiliser. Bien au contraire, cet engagement doit être engendré (ou revitalisé s'il a déjà existé auparavant) et consolidé tout au long du processus.
- L'utilisation d'une évaluation est différente de l'élaboration de rapports et de la diffusion des résultats. Ceux-ci sont des moyens de faciliter l'utilisation, mais ils ne doivent pas être confondus avec des utilisations concrètes telles que la prise de décision, l'amélioration d'un programme ou la production de connaissances.
- Un évaluateur axé sur l'utilisation doit avoir des capacités allant au-delà de l'expertise technique et méthodologique. Pour s'assurer qu'une évaluation n'est pas mise en péril par des processus de groupes destructifs ou des jeux de pouvoir, il doit avoir un certain savoir-faire politique, être habile en animation de groupe, et être un communicateur efficace.
- Une évaluation axée sur l'utilisation peut impliquer du temps et des coûts qui sont loin d'être négligeables. En revanche, les bénéfices qu'on peut retirer d'un tel investissement, soit une utilisation plus efficace de l'évaluation, sont eux aussi loin d'être insignifiants. Ces coûts doivent être explicites dans les budgets d'évaluation, de sorte que le souci de l'utilisation ne soit pas négligé ou délaissé à cause d'un manque de ressources.

Les enjeux de l'évaluation axée sur l'utilisation

L'approche d'une évaluation axée sur l'utilisation comporte un certain nombre d'enjeux associés à la qualité technique, aux changements fréquents des utilisateurs, aux rôles variés de l'évaluateur, à l'utilisation abusive de l'évaluation, et à des questions éthiques.

La participation des utilisateurs et la qualité de l'évaluation

L'implication active des utilisateurs dans le choix des méthodes ne devrait pas mettre en péril la qualité ou la rigueur de l'évaluation. Les utilisateurs d'une évaluation veulent des informations qui sont à la fois valides *et* utiles; la participation des utilisateurs, quelles que soient leurs compétences techniques, accroît l'utilité des résultats. Un évaluateur habile doit pouvoir aider les non-scientifiques à juger d'eux-mêmes les avantages et les désavantages des différentes options méthodologiques afin que ceux-ci puissent choisir celle qui répondra le mieux à leurs besoins.

Il faut aussi reconnaître que la validité et la robustesse des données d'évaluation peuvent varier d'une situation à l'autre. L'enjeu n'est pas de toujours atteindre une norme absolue de qualité scientifique ou méthodologique; il s'agit plutôt de s'assurer que les méthodes et les mesures choisies sont *appropriées* aux besoins de validité d'une situation particulière. L'évaluateur et les utilisateurs doivent déterminer le degré de validité approprié en tenant compte de ce qu'ils prévoient faire avec les données.

Le changement fréquent des utilisateurs

Le talon d'Achille de l'évaluation axée sur l'utilisation est le changement fréquent des principaux utilisateurs. Étant donné que cette approche dépend beaucoup de la participation des utilisateurs attendus, si certains d'entre eux quittent le processus en cours de route, l'utilisation de l'évaluation peut en souffrir. Des personnes qui se joignent à une évaluation tard dans le processus n'arrivent pas nécessairement avec les mêmes attentes que celles qui ont été présentes depuis le début. Le meilleur moyen de contrer les effets de tels changements est de travailler avec un groupe diversifié d'utilisateurs. Ainsi, le départ éventuel d'un ou deux utilisateurs sera moins perturbant. Par contre, en présence de changements importants dans le groupe d'utilisateurs, il peut être nécessaire de redémarrer

le processus en négociant de nouveau le plan d'évaluation et les engagements envers son utilisation.

Il n'est pas toujours facile pour un évaluateur de cibler les bons utilisateurs, de les amener à consacrer de leur temps et de leurs efforts à l'évaluation, de gérer les dynamiques politiques, d'établir une crédibilité et de conduire une évaluation de manière éthique. Tous ces défis sont liés d'une façon ou d'une autre à la relation entre l'évaluateur et les utilisateurs attendus. Quand de nouveaux utilisateurs remplacent ceux qui sont partis, il faut développer de nouvelles relations. Cela peut entraîner des délais dans le processus d'évaluation, mais ils en valent la peine si l'utilisation éventuelle de l'évaluation en est augmentée.

Un rôle qui varie en fonction du but de l'évaluation

Le rôle de l'évaluateur varie selon l'objectif visé et le type d'évaluation. S'il s'agit de générer des connaissances généralisables sur les liens causaux entre une intervention et ses effets, il faut appliquer des méthodes rigoureuses empruntées aux sciences sociales. Dans ce cas, le rôle principal de l'évaluateur est celui d'expert en méthodologie. Quand il faut déterminer la valeur globale d'un programme, l'évaluateur doit jouer un rôle de juge. Si l'évaluation a été commandée afin de répondre à des questions liées à l'imputabilité publique, l'évaluateur assume un rôle de vérificateur, d'inspecteur ou d'enquêteur indépendant. Quand l'amélioration du programme est le but principal de l'exercice, l'évaluateur joue un rôle de conseiller et de facilitateur auprès du personnel du programme. Enfin, en tant que membre d'une équipe de conception de programmes, l'évaluateur peut jouer un rôle plutôt consultatif.

Il existe cependant un rôle que l'évaluateur axé sur l'utilisation doit toujours assumer, et c'est celui de négociateur. Au tout début de l'évaluation, l'évaluateur doit négocier avec les principaux utilisateurs attendus les rôles qu'il aura à jouer. Tous les rôles sont proposés, tout comme toutes les méthodes sont des options possibles. La sélection du rôle de l'évaluateur dépend des utilisations souhaitées de l'évaluation.

L'utilisation abusive des évaluations

Un évaluateur axé sur l'utilisation vise toujours à optimiser l'usage des résultats et du processus, mais il doit aussi porter une attention à la possibilité qu'une évaluation soit utilisée de manière abusive. Alkin et Coyle (1988) ont fait une distinction importante entre une « mauvaise évaluation », dans laquelle l'évaluateur ne fait pas bien son travail ou n'adhère pas aux principes et aux normes de la profession, et une « mauvaise utilisation », par laquelle les utilisateurs manipulent l'évaluation pour tordre les résultats ou corrompre l'enquête. La *mauvaise utilisation* est parfois hors du contrôle de l'évaluateur, mais il a toujours un contrôle direct sur la qualité de son travail, pour éviter une *mauvaise évaluation*.

Les enjeux éthiques

La possibilité qu'un évaluateur axé sur l'utilisation se laisse trop influencer par les différentes parties prenantes d'une évaluation peut constituer une préoccupation pour certains. Comment les évaluateurs peuvent-ils maintenir leur intégrité s'ils sont engagés dans une collaboration étroite avec les décideurs? Comment un évaluateur peut-il prendre en compte les facteurs politiques d'une situation sans qu'il devienne lui-même un outil politique pour une ou plusieurs parties prenantes de l'évaluation?

Les évaluateurs se retrouvent donc face à un dilemme: une relation trop étroite avec les décideurs peut mettre en péril la crédibilité scientifique de l'évaluation, tandis qu'une collaboration trop distante peut entraver son utilisation optimale. Des relations étroites entretenues avec les décideurs soulèvent des enjeux éthiques liés à l'intégrité, l'objectivité et la corruptibilité de l'évaluateur. Le fait de sélectionner les acteurs qui participent activement à l'évaluation comporte des enjeux éthiques liés à l'accès à l'information et au pouvoir associé à ces connaissances. Ces deux enjeux tournent autour d'une question éthique fondamentale: l'évaluation et l'évaluateur sont au service de qui?

Les évaluateurs doivent porter une attention particulière aux intérêts représentés dans l'évaluation et à l'accès aux informations qui découlent de celle-ci. La partie « actif » du « actif-réactif-interactif-adaptatif » signifie que les évaluateurs doivent apporter leurs propres préoccupations, questions et valeurs à la table de négociation. L'évaluateur est lui aussi partie

prenante d'une évaluation puisque sa réputation et sa crédibilité sont en jeu. Le processus « actif-réactif-interactif-adaptatif » implique une obligation, de la part de l'évaluateur, de représenter les principes et les normes de la profession, ainsi que ses propres valeurs et son intégrité, tout en respectant les croyances et les préoccupations des autres utilisateurs principaux.

Un deuxième défi porte sur la façon dont les intérêts des parties prenantes sont représentés dans un processus axé sur l'utilisation. Dans la mesure du possible, elles se représentent eux-mêmes dans les négociations et le processus de l'évaluation. Lorsque les intérêts de personnes défavorisées sont en jeu, il ne faut pas que celles-ci soient représentées par les groupes plus favorisés de manière potentiellement condescendante ou paternaliste. Il faut plutôt explorer les façons de les faire participer activement et équitablement tout au long du processus d'évaluation.

Conclusion

Le principe fondamental de l'évaluation axée sur l'utilisation – travailler avec les principaux utilisateurs attendus pour atteindre l'utilisation souhaitée – a pris une place centrale dans la pratique de la plupart des évaluateurs professionnels. Cousins et ses collègues (1996) ont effectué un sondage auprès de 564 évaluateurs et 68 praticiens, tous membres d'associations professionnelles d'évaluation au Canada et aux États-Unis. Le sondage comportait une liste de croyances possibles avec lesquelles les répondants pouvaient être d'accord ou non. Le plus grand consensus a été obtenu pour l'affirmation « Les évaluateurs devraient formuler des recommandations à partir de l'étude ». La proposition qui a généré le deuxième plus grand consensus (71 %) fut « La principale fonction de l'évaluateur est de maximiser les utilisations attendues des données de l'évaluation par les utilisateurs attendus ». On peut donc dire que depuis la première édition de *Utilization-Focused Evaluation*, il y a trente ans, son principe de base est passé du statut d'idée controversée (Alkin, 1990) à celui d'une philosophie d'évaluation courante.

Pour aller plus loin :

COUSINS, J. B. et L. M. SHULHA (2006), « A Comparative Analysis of Evaluation Utilization and Its Cognate Fields of Inquiry: Current Issues and Trends », dans SHAW, I. F., J. C. GREENE et M. M. MARK (dir.), *The Sage Handbook of Evaluation: Policies, Programs and Practices*, Thousand Oaks, Sage Publications, p. 266-291.

Une excellente revue de l'ensemble de la littérature portant sur l'utilité de l'évaluation, avec un point de vue interdisciplinaire qui inclut les recherches sur l'utilisation des connaissances et plusieurs théories élaborées dans des pratiques connexes.

COUSINS, J. Bradley (dir.) (2008), *Process Use in Theory, Research and Practice: New Directions for Evaluation no 116*, Jossey-Bass / American Evaluation Association.

On retrouve dans cet ouvrage, d'une part, les résultats des recherches les plus récentes sur l'utilisation des connaissances et, d'autre part, les variations conceptuelles en renforcement des capacités en évaluation et en utilisation des résultats des évaluations.

PATTON, M. Q. (2008), *Utilization-Focused Evaluation*, 4^e éd., Thousand Oaks, Sage Publications.

Cette édition a été révisée afin d'inclure les recherches et théories les plus récentes. En plus d'y retrouver des exemples pris à l'échelle internationale, on y approfondit les développements récents des systèmes de pensée et de la science de la complexité dans le domaine de l'évaluation.

PATTON, M. Q. (2012), *Essentials of Utilization-Focused Evaluation*, Thousand Oaks, Sage Publications.

Basé sur l'ouvrage *Utilization-Focused Evaluation*, ce livre, plus synthétique et condensé, présente un cadre général et une liste révisée d'étapes essentielles pour concevoir et mener des évaluations axées sur l'utilisation.

Internet

Centers for Disease Control: *CDC Evaluation Working Group*.

Excellentes ressources en ligne pour l'évaluation de la santé publique, qui donnent accès aux méthodes et à l'éthique, en plus d'offrir des liens vers d'autres sites pertinents: www.cdc.gov/eval/resources.htm#logic

Resources for Methods in Evaluation and Social Research.

Recherches de base et sur l'évaluation; liens vers d'autres sources: <http://gsociology.icaap.org/methods/>

9

Une stratégie pour composer avec les contraintes inhérentes à la pratique

Michael Bamberger et Jim Rugh

La plupart des évaluateurs ont connu ces situations où l'évaluation arrive bien tard dans la vie des programmes : ceux-ci sont en place, voire presque arrivés à leur terme, avant même que les organismes d'exécution ou de financement ne commencent à penser sérieusement à évaluer dans quelle mesure ils atteignent leurs objectifs et produisent les effets escomptés. Généralement, cet intérêt tardif pour l'évaluation est motivé par la nécessité d'avoir des preuves solides sur lesquelles se baser pour prendre la décision de maintenir le programme, ou encore de l'implanter ailleurs. Lorsque les évaluations débutent enfin, c'est généralement avec des contraintes budgétaires et temporelles, ou encore avec un accès limité aux données de base ou à des groupes témoins. Dans ce contexte, il est parfois difficile, sinon impossible, d'utiliser les modèles d'évaluation recommandés dans les manuels pour leur rigueur méthodologique.

En 2002, lorsque les auteurs du présent chapitre ont commencé à organiser des ateliers sur les évaluations en contexte de contraintes de budget, de temps, de données et politiques, ils pensaient que les évaluateurs des pays en développement se montreraient les plus intéressés. Or, bien que

plus de ressources soient affectées à l'évaluation des grands programmes gouvernementaux des pays comme le Canada, les États-Unis et les pays européens, de nombreux évaluateurs de ces contrées ont dit devoir composer avec des contraintes similaires. Par ailleurs, dans les pays industrialisés comme dans les pays en développement, la plupart des évaluations sont mises en œuvre dans des contextes politiques où les organismes de financement, les clients et les principaux acteurs des programmes ont des opinions bien arrêtées sur ce qui constitue de « bonnes méthodes » d'évaluation, sur les types et quantités d'informations à collecter, et sur les groupes auprès desquels on devrait ou non recueillir une opinion sur les conclusions de l'évaluation. Et s'ils devraient y avoir accès ou non. Les évaluateurs en début de carrière découvrent rapidement que les questions « techniques » – telles que le recours à des groupes témoins sélectionnés aléatoirement, le choix entre des modèles qualitatifs, quantitatifs ou multiméthodes (voir chapitre 7), ou encore la sélection des personnes à interviewer et des questions à poser – peuvent déclencher de vives réactions chez les clients et acteurs impliqués. Mais souvent, ces réactions n'auront rien à voir avec des raisons techniques, comme d'éviter d'interviewer les gens qui critiquent le programme ou de se soustraire à la controverse politique en n'organisant aucun entretien avec des membres des groupes de comparaison qui pourraient trouver injuste de ne pas bénéficier des retombées positives du programme. Les évaluateurs qui travaillent en milieu universitaire s'aperçoivent très rapidement que certains paradigmes de recherche et d'évaluation sont acceptés, alors que d'autres ne le sont pas. Récemment, lors d'une séance de travail consacrée à l'évaluation multiméthodes offerte par l'Association américaine d'évaluation, des étudiants diplômés d'une université canadienne ont dit avoir de la difficulté à convaincre leurs directeurs de thèse qu'un modèle d'évaluation multiméthodes pourrait être approprié pour leur recherche doctorale.

En dépit des circonstances complexes qui influencent la réalisation de nombreuses évaluations, la demande d'évaluations systématiques de la part d'organismes, de gouvernements, de la société civile et de bénéficiaires de programmes ne cesse d'augmenter. On attend de ces évaluations qu'elles mesurent si les programmes produisent les résultats escomptés et qu'elles indiquent s'ils pourraient ou devraient être poursuivis ou reproduits, et à qui ils ont profité ou non. On réalise également de plus en plus que les conclusions des évaluations doivent reposer sur des preuves solides,

pas seulement sur des opinions, et ce, même si le débat continue de faire rage quant à la définition d'une *preuve solide*.

Bien qu'il n'existe aucune donnée précise à ce sujet, nous estimons que la majorité des évaluations de projets et de programmes de développement qui reçoivent un financement international ne satisfont pas aux exigences de base d'un modèle d'évaluation quantitative rigoureux du point de vue méthodologique. La plupart de ces évaluations commencent tard dans le cycle de vie du programme et les évaluateurs n'ont pas accès à des données de base systématiques, ou alors aucun groupe de comparaison n'est inclus. Pour ne citer qu'un exemple de cela, une méta-évaluation de 67 évaluations menées par CARE International sur ses projets d'aide d'urgence et de développement dans divers pays a révélé que, pour environ 75 % des évaluations pour lesquelles l'information était disponible, le modèle utilisé ne comportait une collecte de données effectuée qu'après que le projet eut été mis en œuvre depuis un certain temps, et ce, généralement uniquement auprès des communautés visées par le projet. Autrement dit, ces évaluations ne comprenaient ni données de base ni données sur un groupe de comparaison. Et il semble que cela soit assez typique des évaluations faites par d'autres organismes. En fait, les évaluations n'impliquent bien souvent qu'une trop brève visite sur le site du projet.

Une des raisons pour lesquelles Bamberger et ses collègues ont rédigé *RealWorld Evaluation* (2006) est que, outre le fait que les contraintes rencontrées dans la réalité entraînent l'impossibilité, pour la majorité des évaluateurs, d'utiliser un modèle d'évaluation un tant soit peu rigoureux, très peu de manuels fournissent des pistes sur la façon d'effectuer des évaluations dans ces conditions. Si, dans la plupart des livres, on parle de la dimension politique des évaluations, il est rare qu'on y traite aussi des stratégies pour composer avec des contraintes budgétaires ou temporelles, ou encore avec l'absence de données de base. Qui plus est, très peu de ces ouvrages donnent des instructions systématiques sur les procédures à suivre pour déterminer le degré de validité des divers modèles d'évaluation en présence de ces contraintes, et sur les moyens de surmonter ces obstacles une fois qu'ils ont été identifiés.

De nombreux évaluateurs se retrouvent donc livrés à eux-mêmes et finissent souvent par utiliser des méthodologies d'évaluation qui manquent de rigueur, sans que cela soit nécessaire. Dans de telles situations, les contraintes budgétaires et temporelles sont fréquemment utilisées

comme excuse (« Il est injuste de me demander si les données présentées dans mon rapport sont représentatives, alors que je n'avais que deux semaines pour travailler sur le terrain »). L'expression « bâclée », utilisée pour qualifier une évaluation, revient beaucoup trop souvent...

L'approche de l'évaluation en situation réelle a donc été spécifiquement conçue afin de répondre aux besoins de stratégies pratiques pour parvenir à la plus grande rigueur possible dans les contextes particuliers dans lesquels beaucoup d'évaluations doivent être effectuées.

L'approche de l'évaluation en situation réelle

L'approche de l'évaluation en situation réelle (ESR) a été créée pour aider les nombreux évaluateurs, tant dans les pays en développement que dans les pays industrialisés, qui doivent mener des évaluations en composant avec des contraintes associées au financement, aux délais, à la disponibilité des données et au contexte politique. De plus, il arrive couramment que le client ne commande une évaluation que tard dans le projet, lorsqu'il réalise qu'il aura besoin des conclusions d'une évaluation dans le processus de révision du projet pour déterminer s'il doit continuer à le soutenir ou s'il devrait lancer une deuxième phase de plus grande envergure. Parfois, quand approche le moment de prendre une décision à ce propos, l'organisme de financement découvre subitement qu'il ne dispose pas d'informations solides pour déterminer s'il financera ou non le projet à l'avenir. Il arrive aussi que l'organisme d'exécution se rende compte qu'il ne possède pas les preuves nécessaires pour démontrer que le projet atteint bien ses objectifs. L'évaluateur appelé à ce stade se fera quelquefois dire qu'il est impératif d'effectuer son évaluation avant une certaine date, et qu'elle doit produire des résultats « rigoureux » quant à l'impact du projet, même si, malheureusement, les fonds disponibles demeurent très limités.

Un autre scénario est possible. L'évaluateur est appelé suffisamment tôt, mais il découvre que, pour des raisons budgétaires, politiques ou méthodologiques, il ne pourra pas collecter de données sur un groupe de comparaison afin de mieux déterminer l'impact du programme en comparant des participants à des non-participants. Ou encore, il s'avère impossible de collecter des données de base sur la population visée par le projet pour analyser ses progrès au fil du temps ou l'impact qu'a eu le

projet sur elle. Des contraintes de disponibilité des données peuvent aussi être causées par la difficulté de collecter des informations sur des sujets délicats comme le VIH/sida, la violence familiale, la reconstruction à la suite d'un conflit ou les activités illégales telles que le travail du sexe, le trafic de narcotiques ou la corruption politique.

Déterminer le modèle d'évaluation qui convient le mieux dans de telles conditions peut se révéler un exercice ardu nécessitant des compromis entre les ressources disponibles et les pratiques acceptables en matière d'évaluation. D'ordinaire, les clients se soucient plus du budget et de l'échéancier, laissant les principes de base du métier passer au second plan. Et si aucun équilibre satisfaisant n'est trouvé entre ces éléments, un problème dont beaucoup se plaignent peut alors survenir : l'utilisation limitée des résultats de l'évaluation (voir Patton, 1997 ; Banque mondiale, 2004 et 2005). L'ESR est donc une approche qui propose des solutions afin de surmonter les difficultés bien réelles que l'on rencontre dans l'univers de la pratique de la discipline de l'évaluation.

De plus, du fait des difficultés occasionnées par toutes ces contraintes, on ne se soucie guère, dans la majorité des cas, d'opter pour une méthodologie rigoureuse ou de veiller à identifier et à surmonter les obstacles à la validité des résultats. L'ESR vise précisément à réduire ces risques potentiels et repose sur une approche en sept étapes, schématisées dans la figure 9.1 et détaillées ci-après.

1. *Cerner le champ de l'évaluation.* Comprendre les besoins d'information du client et le contexte dans lequel s'inscrit l'évaluation ; définir une théorie de programme qui décrit les objectifs et le fonctionnement prévus du programme ; identifier les contraintes (reliées au budget, au temps ou à la disponibilité des données) et toutes les influences politiques qui agissent sur les façons dont l'évaluation sera mise en œuvre, et sur la diffusion et l'utilisation de ses résultats.
2. *Composer avec les contraintes budgétaires (stratégies).* Réduire les frais en simplifiant le modèle d'évaluation, en réduisant la quantité de données à collecter en faisant un meilleur usage des consultants onéreux et en rationalisant l'analyse des données.
3. *Composer avec les contraintes temporelles (stratégies).* Méthodes de collecte de données rapide ; planification adaptée pour éviter les retards et les goulots d'étranglement, en particulier pendant les brèves

périodes où les consultants externes interviennent; recours à des vidéoconférences.

4. *Composer avec les contraintes en termes de données (stratégies).* Reconstitution des données de base ou des données du groupe de comparaison si elles n'ont pas été collectées; utilisation judicieuse des données secondaires; importance accrue accordée à la collecte de données auprès des groupes appropriés (par exemple, les femmes tout autant que les hommes, les minorités, les personnes sans terre et les immigrants clandestins); stratégies permettant de collecter des données sur des sujets délicats et de trouver et d'interviewer les groupes difficiles à joindre.
5. *Composer avec les influences politiques agissant sur la conception et la mise en œuvre de l'évaluation, ainsi que sur la diffusion et l'utilisation de ses résultats.* Identifier les principaux acteurs impliqués et leur perspective politique de l'évaluation, y compris leurs préjugés favorables pour certaines méthodes et même pour certaines conclusions; comprendre comment ces éléments peuvent orienter l'évaluation; élaborer des stratégies pour composer avec la situation politique sans compromettre la rigueur de l'évaluation.
6. *Renforcer le modèle d'évaluation et la validité des conclusions.* Utiliser la liste de contrôle (voir l'annexe du présent chapitre) pour identifier les obstacles à la validité et à la pertinence des conclusions de l'évaluation; élaborer des stratégies pour surmonter ces obstacles et pour améliorer la qualité de l'évaluation.
7. *Aider les clients à tirer parti de l'évaluation.* Veiller à ce que les clients participent activement à l'évaluation dès le début et à ce qu'ils se l'approprient; rester en contact avec le client tout au long de l'évaluation pour s'assurer que les principaux rapports, une fois rendus publics, ne contiennent pas de surprises majeures pour le client; adapter la présentation des conclusions en fonction des habitudes de communication des différents acteurs concernés.

Si l'ESR ne propose pas de nouvelles méthodes de collecte ou d'analyse des données, l'approche contribue à plusieurs égards à la réalisation d'évaluations qui composent avec les contraintes rencontrées dans la réalité.

FIGURE 9.1**L'approche de l'évaluation en situation réelle**

Premièrement, l'ESR suggère des façons de puiser dans un large éventail d'approches et de méthodes d'évaluation pour composer avec les quatre grands types de contrainte. On insiste en permanence sur le recours systématique à des approches multiméthodes qui combinent les approches quantitatives et qualitatives (chapitre 8). Ces dernières sont jugées essentielles pour plusieurs raisons. Elles offrent d'abord à l'évaluateur la possibilité de se servir du plus grand nombre possible de méthodes et d'instruments d'évaluation; elles mènent à des conclusions plus valides en fournissant au moins deux estimations sur les principaux indicateurs (triangulation); elles permettent une analyse et une interprétation plus approfondies et plus riches du contexte dans lequel fonctionne le programme et contribuent également à rendre la collecte de données plus abordable et plus rapide.

Deuxièmement, avec l'ESR, l'accent est davantage mis sur l'importance d'orienter l'évaluation vers un modèle théorique (logique) de programme. En présence de contraintes budgétaires et temporelles, un tel modèle peut aider à identifier les postulats de base, les liens et les hypothèses du programme qui devraient être vérifiés. À cet égard, il est particulièrement utile de procéder à une analyse de la chaîne des résultats. De même, *l'analyse contextuelle* est une composante essentielle du modèle théorique de programme de l'ESR. Elle comprend l'analyse de l'influence des conditions politiques, économiques, institutionnelles et socioculturelles sur la mise en œuvre du projet et sur ses effets. Elle peut conduire à une analyse qualitative et descriptive, ou encore les facteurs peuvent être convertis en échelles de valeurs et être incorporés à l'analyse multivariée. Le modèle théorique fait également partie de l'analyse des processus pour déterminer ce qui s'est véritablement produit durant l'implantation du projet. En effet, l'analyse des *échecs au moment de l'implantation* est souvent cruciale afin d'expliquer pourquoi un projet n'a pas eu les impacts escomptés.

Troisièmement, sept des modèles d'évaluation d'impacts les plus courants sont décrits dans l'ESR, ainsi que des scénarios dans lesquels chacun de ces modèles peut être utilisé. Ces derniers vont des modèles expérimentaux et quasi expérimentaux les plus rigoureux (avec des mesures avant/après et des groupes d'intervention et de contrôle) aux modèles non expérimentaux dans lesquels il n'y a ni groupe de comparaison ni données de base sur le

groupe d'intervention, en passant par des modèles quasi expérimentaux plus faibles, dans lesquels un ou plusieurs points de collecte de données sont supprimés (chapitre 20). Enfin, les obstacles à la validité sont aussi passés en revue pour chacun de ces modèles (voir tableau 9.1).

Quatrièmement, certaines techniques sont proposées en vue de reconstituer l'information sur la situation des groupes d'intervention et de comparaison avant le début du projet (données de base rétrospectives). Celles-ci comprennent l'utilisation exhaustive de données secondaires, la collaboration avec les organismes pour rendre les données de suivi régulières du projet plus utiles à l'évaluation, les techniques d'évaluation participative telles que les méthodes d'évaluation participative rapide et le recours aux souvenirs personnels. Bien que les problèmes de fiabilité en matière de recours aux souvenirs soient reconnus, un certain nombre de techniques permettant d'y remédier sont décrites dans l'ESR.

Cinquièmement, l'approche de l'ESR pour le contrôle de la qualité en sept étapes propose des mesures correctives qui peuvent être employées à différents stades du processus d'évaluation – certaines après qu'une ébauche du rapport d'évaluation a été produite – et qui font en sorte que la qualité optimale de bien des évaluations quantitatives repose sur un modèle ayant un prétest, un post-test ainsi qu'un groupe de comparaison pour évaluer les changements ou impacts produits par le projet ou le programme.

Sixièmement, bien des évaluations quantitatives reposent sur un modèle ayant un prétest/post-test et un groupe de comparaison pour évaluer les changements ou les impacts produits par le projet ou le programme. Mais cette approche, lorsque utilisée de manière isolée, comporte deux limites importantes. D'une part, elle ne tient pas compte des particularités du contexte socioéconomique et politique unique qui affecte chaque projet. D'autre part, elle accepte implicitement l'idée que chaque projet a été implanté tel que prévu et de façon uniforme d'un site à l'autre. Aussi, une des contributions de l'ESR est de souligner la nécessité de regarder dans la « boîte noire » du processus d'implantation du projet pour voir ce qui s'est réellement produit lors de la mise en œuvre et pour déterminer les variations à ce propos entre les différents sites du projet. La qualité de

la mise en œuvre doit également être évaluée. Cette contribution de l'ESR est majeure, car dans la réalité, il arrive très souvent que certaines caractéristiques d'un programme ne soient pas mises en œuvre, ou qu'elles le soient si mal que les impacts escomptés ne peuvent pas s'être concrétisés. D'autres fois, ces impacts semblent atteints, mais ce qui s'est vraiment déroulé dans le cadre du projet a été bien différent de ce qui avait été initialement prévu!

TABLEAU 9.1

**Forces et faiblesses des sept modèles
quasi expérimentaux les plus utilisés**

Modèle	Avantages	Inconvénients
1. Modèle longitudinal complet, avec observations pré-, à mi-parcours, post- et ex-post et avec un groupe de comparaison.	Le modèle le plus rigoureux. Peut être nécessaire pour mettre à l'épreuve un nouveau projet qui, si son impact est démontré, pourra être mis en œuvre à plus grande échelle.	Modèle le plus onéreux et celui qui demande le plus de temps, et donc le plus difficile à mettre en œuvre.
2. Modèle avec mesure avant et après le projet et avec un groupe de comparaison.	Parmi les méthodes quasi expérimentales polyvalentes, c'est le modèle le plus rigoureux. Avec un groupe témoin bien choisi, il fournit de bonnes estimations des impacts du projet.	Suppose que le groupe de comparaison est raisonnablement similaire au groupe de projet et d'accord pour participer à deux sondages, même s'il n'a rien à y gagner.
3. Modèle longitudinal tronqué avec mesures avant et après le projet et avec un groupe de comparaison.	• Permet l'observation des processus ainsi que des impacts. • Modèle raisonnablement rigoureux, en particulier pour les projets où la mise en œuvre démarre lentement et où l'on ne manque pas grand-chose en commençant plus tard.	• N'évalue pas la mise en œuvre du projet. • Ne commence pas avant la mi-parcours, et n'évalue donc pas la mise en route du projet et le début de sa mise en œuvre.
4. Modèle avec mesures avant et après le projet, combiné avec une analyse post-test du groupe de projet et du groupe de comparaison.	• Permet de déterminer si le modèle de projet produit les extrants visés. • Permet d'identifier les similarités et les différences avec des domaines similaires. • Permet de déterminer dans quelle mesure le projet pourrait être répliqué.	• Ne permet pas de déterminer si les différences observées à la fin du projet avec le groupe de comparaison sont attribuables au projet ou à des différences préexistantes entre les deux groupes. • Ne permet pas de contrôler si des événements de l'histoire locale pourraient influencer les effets.

Modèle	Avantages	Inconvénients
5. Modèle avec mesure après le projet et avec un groupe de projet et un groupe de comparaison.	Permet d'évaluer des projets pour lesquels on a mis en œuvre des interventions soigneusement testées ou qui se déroulent dans des endroits isolés où d'autres interventions externes n'interfèrent pas.	• Ne permet pas d'estimer avec exactitude les impacts du projet. • Ne permet pas de contrôler le rôle de l'histoire locale. • Ne permet pas de déterminer les possibilités de reproduire le projet à plus grande échelle. • Ne permet pas d'étudier le processus de mise en œuvre du projet.
6. Modèle avec mesure avant et après le projet sans groupe de comparaison.	Donne une estimation approximative des impacts du projet.	• Ne permet pas d'estimer avec exactitude les impacts du projet. • Difficile d'établir que les changements observés sont dus au projet et non à d'autres facteurs ou interventions. • Ne permet pas la comparaison avec d'autres collectivités. • Ne permet pas de contrôler si d'autres variables ont joué un rôle au moyen d'une analyse multivariable.
7. Modèle avec mesure après le projet sans groupe de comparaison.	• Utile pour les études préliminaires lorsque l'on veut avoir une idée générale de l'efficacité du projet. • Donne une première estimation approximative des résultats, en particulier pour des projets petits ou isolés.	• Ne permet pas d'obtenir des estimations raisonnablement précises de l'impact du projet. • Difficile d'être sûr que les changements observés sont dus au projet et non à d'autres facteurs ou interventions. • Ne permet pas de contrôler le rôle d'événements extérieurs. • Ne permet pas d'obtenir des données comparatives pour estimer les possibilités de répliquer le projet.

Note : Pour les modèles 1 à 5, qui utilisent des groupes de comparaison, une analyse beaucoup plus rigoureuse peut être faite en utilisant une analyse par régression multiple pour contrôler statistiquement les différences sur le plan des caractéristiques du groupe de projet et du groupe de comparaison. Lorsque des données secondaires fiables sont disponibles, ces modèles peuvent en outre être renforcés au moyen de techniques statistiques d'appariement, sur des scores de propension ou des variables instrumentales par exemple.

L'ESR et les quatre types de contraintes

On trouvera ci-dessous une description plus détaillée des quatre types de contraintes auxquelles sont le plus souvent confrontés les évaluateurs dans la réalité. Il est à noter que les évaluateurs se heurtent généralement à plus d'un type de contraintes à la fois.

Les contraintes budgétaires

Dans certains cas, les fonds nécessaires à l'évaluation n'ont pas été inclus dans le budget initial du projet, ce qui implique qu'elle doit être menée avec un budget inférieur à ce qu'elle nécessiterait normalement. Dans d'autres cas, le budget est réduit en cours d'évaluation alors qu'on ne s'y attendait pas, ce qui entraîne parfois l'impossibilité de collecter les données souhaitées ou de reconstituer les données de base ou celles du groupe de comparaison. Le manque de fonds peut aussi créer ou exacerber des contraintes temporelles, car les chercheurs ne peuvent pas toujours passer sur le terrain autant de temps qu'ils le jugent nécessaire.

Les contraintes temporelles

La contrainte temporelle la plus fréquente se produit lorsque l'évaluateur n'est appelé qu'une fois le projet déjà bien avancé. Il ne dispose alors que d'une période jugée beaucoup trop courte pour faire l'évaluation, tant en termes de perspective longitudinale durant la vie du projet qu'en ce qui concerne le temps alloué pour effectuer l'évaluation de fin de projet. De plus, les contraintes temporelles rendent souvent impossible l'utilisation d'un modèle d'évaluation prétest/post-test avec une mesure des données de base pouvant être répétée, une fois le projet mis en œuvre. Le temps disponible pour planifier les consultations auprès des acteurs concernés, les visites sur les lieux et le travail sur le terrain ainsi que l'analyse des données doivent parfois être considérablement abrégés pour pouvoir remettre le rapport à la date prévue. Ces pressions temporelles sont particulièrement problématiques pour un évaluateur qui ne connaît pas bien la région ou le pays où a lieu l'évaluation, ou qui n'a pas le temps de se familiariser avec les communautés et organismes concernés par l'étude et de bâtir des rapports de confiance avec eux.

Les contraintes en termes de disponibilité des données

Lorsque l'évaluation ne commence que tard dans le cycle d'un projet, les données de base sur la situation du groupe ciblé avant le commencement du projet sont limitées ou inutilisables, voire inexistantes. Même si les dossiers du projet sont accessibles, le plus souvent ils ne sont pas organisés de manière à permettre une analyse comparative pré-post. De plus, les dossiers de projet et autres documents présentent souvent un biais de mémoire systématique ou témoignent d'un manque de rigueur dans leur gestion. En outre, si des données secondaires sont disponibles pour une période voisine de la date de démarrage du projet, il est rare qu'elles correspondent exactement aux populations visées par ce projet. Par ailleurs, il est souvent difficile de recruter des participants pour les groupes de comparaison, car ils ne retirent aucun avantage de leur participation et n'ont aucun intérêt dans l'intervention. De ce fait, bien des projets pour lesquels on avait initialement prévu d'utiliser un groupe de comparaison pour l'évaluation ne réussissent pas à rassembler un nombre suffisant de gens pour remplir cette fonction.

Les influences politiques

Nous utilisons les expressions de contraintes ou d'influences *politiques* au sens large pour évoquer non seulement les pressions exercées par les organismes gouvernementaux, mais aussi les exigences des organismes de financement ou de réglementation, les revendications provenant des acteurs concernés et, au sein même de l'équipe d'évaluation, les divergences d'opinions quant aux cadres conceptuels et aux méthodes de collecte de données. Les évaluations sont souvent menées dans des contextes où les enjeux politiques et éthiques influent sur le choix du modèle d'évaluation et sur l'utilisation des résultats. Tous les programmes touchent une partie de la population, et la majorité d'entre eux utilisent des fonds publics toujours limités et souvent rares. Les décisions fondées sur les résultats de l'évaluation accroissent parfois la compétition pour obtenir des fonds et peuvent entraîner l'expansion ou la suppression de programmes. Elles peuvent également servir les intentions ou le programme d'un groupe politique.

Conclusion

L'approche de l'ESR peut être utile à différents stades d'un projet ou d'un programme : au début, pendant la phase de planification, en cours de mise en œuvre ou à la fin. Lorsque le processus de planification de l'évaluation commence au début du projet, l'ESR est utilisée pour identifier différentes options qui rendent les collectes de données initiales et subséquentes plus économiques et plus rapides. Elle permet également de tirer le meilleur parti possible des données disponibles et de mieux comprendre les besoins du client et le contexte dans lequel sera effectué l'exercice. Si l'évaluation ne commence pas avant que le projet ne soit déjà en cours, l'ESR servira à identifier et à jauger les différents modèles d'évaluation auxquels on peut avoir recours, compte tenu des contraintes budgétaires et temporelles, et à trouver des façons de reconstituer les données de base. Aussi, on prendra soin d'évaluer les forces et faiblesses des données de surveillance du projet qui sont disponibles et de déterminer l'accessibilité et la qualité des données secondaires issues d'autres sources. De même, on pourra examiner la possibilité de créer un groupe de comparaison. Enfin, si l'évaluation ne commence que vers la fin du projet (ou une fois que le projet est déjà terminé), l'ESR aura la même utilité que dans la situation précédente, avec cependant un choix restreint de modèles d'évaluation utilisables.

Quel que soit le stade du programme durant lequel l'approche de l'ESR est utilisée, un de ses aspects novateurs consiste à suggérer des mesures qui, en cas de contraintes majeures, peuvent être prises pour renforcer la crédibilité et la validité des conclusions d'une évaluation potentiellement moins rigoureuse. Une fois que les personnes chargées de l'évaluation auront choisi les questions auxquelles il faut répondre, elles devront choisir les méthodes les plus appropriées, réalistes et rigoureuses en fonction des buts poursuivis par l'évaluation et du contexte dans lequel elle prend place.

Le débat continue, chez les professionnels de la discipline, pour déterminer si les essais contrôlés randomisés (ECR) devraient être acceptés comme la « norme idéale » pour l'évaluation d'impacts. L'avantage méthodologique des ECR est qu'ils éliminent les biais de sélection et les biais d'échantillonnage, deux des plus grandes menaces à la validité des résultats (même ceux qui sont obtenus avec des modèles quasi expérimentaux rigoureux). Cependant, un grand nombre d'évaluateurs professionnels

rejettent l'affirmation voulant que les modèles quantitatifs rigoureux soient les meilleurs moyens, ou même les seuls moyens, d'évaluer les impacts d'un projet ou d'un programme. Si les avis diffèrent quant aux mérites des ECR, le débat a toutefois aidé les évaluateurs à canaliser leurs efforts sur la nécessité d'accroître la rigueur de nombreuses évaluations et sur l'importance de l'analyse d'une situation hypothétique en l'absence d'intervention (*counterfactual*) qui permette d'établir quelle aurait été la situation de la population visée par le projet s'il n'avait pas eu lieu.

Malheureusement, la littérature dans le domaine de l'évaluation mentionne peu de façons de renforcer la rigueur méthodologique des modèles d'évaluation qui ne peuvent en aucun cas s'approcher des modèles recommandés dans la plupart des manuels scientifiques. L'approche de l'évaluation en situation réelle est donc conçue pour combler cette lacune entre théorie et pratique en proposant des stratégies pour s'adapter aux contraintes de la réalité, et cela dans l'objectif de produire des modèles hautement rigoureux compte tenu des conditions réelles fréquemment rencontrées en évaluation. L'ESR fournit également une liste de contrôle qui permet d'identifier les obstacles à la validité des résultats et suggère des façons de les surmonter ou d'en minimiser l'importance.

Recourir exclusivement à des ECR ou à des modèles quasi expérimentaux rigoureux comporte aussi des limites. En effet, lorsque aucun impact significatif n'est détecté, d'un point de vue statistique, au terme d'un projet, il n'est pas possible de déterminer *pourquoi* et de conseiller aux décideurs et aux planificateurs de rejeter le modèle de projet (problème de conception) ou de tester le projet à nouveau en faisant davantage attention à la façon dont il est mis en œuvre (problème de mise en œuvre). L'utilité de tels modèles de recherche est donc considérablement restreinte pour les planificateurs et décideurs. L'ESR – axée sur les modèles logiques, l'analyse des processus et la combinaison de méthodes différentes – apporte des indications quant à la difficulté d'obtenir les impacts souhaités et quant aux implications des politiques.

Il n'est pas toujours possible ou nécessaire, pour un évaluateur de l'extérieur ou pour une équipe d'évaluateurs, de collecter des données de base au moment de l'évaluation initiale ou à mi-parcours, ou encore de se servir d'un groupe témoin ou d'un groupe de comparaison. Il existe en effet d'autres façons d'obtenir des preuves, même pour la situation de référence en l'absence d'intervention. L'ESR vise à encourager les éva-

luateurs à diversifier le contenu de leur boîte à outils et à les aider à déterminer quelle combinaison d'outils utiliser, et à quel moment, afin de composer avec les contraintes et d'améliorer la qualité, la crédibilité et l'utilité d'une évaluation.

Pour aller plus loin :

BAMBERGER, M., J. RUGH et L. MABRY (2006), *L'évaluation en situation réelle* « en bref ».

Adapté du chapitre 16 de *RealWorld Evaluation*, des mêmes auteurs, ce document traduit en français avec la coopération de l'UNICEF examine les diverses façons d'appliquer l'approche de l'ESR. Il inclut également certains graphiques et tableaux importants des autres chapitres, dont la version complète de la Liste de contrôle intégrée pour évaluer l'utilité et la validité des modèles qualitatifs, quantitatifs et multi-méthodes. Disponible en ligne sur le site du livre : www.realworldevaluation.org.

BANQUE MONDIALE, Département des évaluations des opérations (2004), *Évaluations marquantes. Des évaluations qui ont contribué à améliorer les résultats et l'impact des programmes de développement*, Washington, Banque mondiale.

L'ouvrage traite des facteurs qui influencent l'utilisation des évaluations de programmes en développement international.

WORLD BANK, Operations Evaluation Department (2005), *Influential Evaluations: Detailed Case Studies*, Washington, World Bank.

Une version anglaise plus détaillée des études de cas de l'ouvrage mentionné ci-dessus.

ANNEXE

Liste de contrôle intégrée pour évaluer l'utilité et la validité des modèles qualitatifs, quantitatifs et multi-méthodes

A. Preuves à l'appui (et objectivité)

Les conclusions sont-elles tirées des preuves dont on dispose et la recherche est-elle relativement dépourvue de préjugés ?

Exemple : *Des hypothèses concurrentes ou rivales ont-elles été envisagées ?*

B. Fiabilité et constance

Le processus de l'étude est-il homogène, cohérent et raisonnablement stable dans le temps, d'un chercheur à l'autre et d'une méthode à l'autre ? Si des modèles émergents sont utilisés, est-ce que les processus qui font évoluer le modèle sont clairement documentés ?

Exemple : *Les données ont-elles été collectées dans un éventail large et bien choisi d'endroits, de moments, de répondants, etc. ?*

C. Validité interne, crédibilité et authenticité

Les constatations sont-elles crédibles pour les personnes sur qui portait l'étude et pour les lecteurs, et a-t-on un portrait fidèle de ce que l'on étudie ?

Exemple : *La triangulation entre les sources de données et les méthodes complémentaires produit-elle en général des conclusions convergentes ? Lorsque les interprétations divergent, les différences sur le plan des interprétations et des conclusions sont-elles notées et discutées ?*

D. Validité externe, transférabilité et adéquation

Les conclusions sont-elles applicables à d'autres contextes et dans quelle mesure peuvent-elles être généralisées ?

Exemple : *Les caractéristiques de l'échantillon de gens, situations, processus, etc. sont-elles décrites de façon suffisamment détaillée pour permettre les comparaisons avec d'autres échantillons ?*

E. Utilisation, application, accent mis sur l'action

Dans quelle mesure les constatations ont-elles été utiles pour les clients, les chercheurs et les collectivités étudiées ?

Exemple : *Les constatations incluent-elles des conseils pour des interventions futures ?*

F. Obstacles à la validité statistique des conclusions

Pour quelles raisons les inférences quant à la covariation entre deux variables peuvent être inexactes ?

Exemple : *Échantillon trop petit pour détecter les effets du programme (efficacité statistique faible), dérogations aux hypothèses des tests statistiques.*

G. Obstacles à la validité interne

Pour quelles raisons les inférences qu'il y a des rapports de cause à effet entre deux variables sont parfois inexactes ?

Exemple : *Biais attribuable à la sélection : Les participants au projet sont souvent différents de ceux du groupe de comparaison, soit parce qu'ils sont auto-sélectionnés, soit parce qu'on a sélectionné pour le projet des gens ayant des caractéristiques particulières (les fermiers les plus pauvres ou les collectivités les mieux organisées).*

H. Obstacles à la validité conceptuelle

Pour quelles raisons les inférences quant aux concepts qui caractérisent la façon dont l'étude est menée sont peut-être inexactes ?

Exemple : *Les indicateurs utilisés ne permettent pas de décrire et de mesurer correctement le concept. Il arrive qu'ils ne mesurent qu'une seule dimension d'un concept complexe ou qu'ils soient mal adaptés pour un contexte culturel particulier.*

I. Obstacles à la validité externe

Pour quelles raisons les inférences quant à la façon dont les résultats de l'étude tiendraient le coup en dépit des variations dans les gens, les contextes, les projets et les effets ne sont pas forcément exactes ?

Exemple : *L'échantillon ne couvre pas l'ensemble de la population à laquelle on s'intéresse.*

Sources : Les sections A à E sont adaptées de Miles et Huberman (1994, chapitre 10, section C). Voir aussi Guba et Lincoln (1989). Les sections F à I sont adaptées de Shadish et al. (2002), tableaux 2.2, 2.4, 3.1 et 3.2, respectivement.

Pour une liste de contrôle intégrée plus complète, voir *RealWorld Evaluation* (Bamberger et al., 2006).

10

Théorie et pratiques des études de cas en évaluation de programmes

Robert K. Yin et Valéry Ridde

Dans le domaine de l'évaluation des programmes, les études de cas peuvent être employées de deux différentes façons, soit elles font partie de l'une des méthodes auxquelles ont recours les évaluateurs dans le processus d'évaluation (par exemple, Cronbach *et al.*, 1980, p. 222-223 ; Datta, 1997, p. 348), soit elles sont utilisées à titre de démarche méthodologique principale (par exemple, Yin, 2000a). La première utilisation est la plus répandue, alors que la seconde pose davantage de difficultés.

L'étude de cas en tant que méthode de recherche¹

Que l'étude de cas soit employée à titre de méthode principale ou secondaire, sa nature demeure la même. Elle consiste dans l'étude approfondie d'un cas dans son contexte. Contrairement aux enquêtes ainsi qu'aux études expérimentales et quasi expérimentales, les études de cas servent à analyser des cas en profondeur et à examiner les relations entre les cas à l'étude et certains éléments contextuels. Pour saisir la complexité d'un cas et de son contexte, les études de cas utilisent plusieurs types de sources de données, tels que les entretiens, les documents, l'observation sur le

1. Il s'agit d'une présentation abrégée des études de cas (voir Yin, 2009, pour une présentation détaillée).

terrain, les archives et l'observation participante. La triangulation des données provenant de ces différentes sources aide à confirmer et à corroborer les résultats. Les données collectées peuvent être de nature qualitative, quantitative ou les deux.

L'étude de cas peut également s'appuyer sur une théorie provisoire. La théorie peut être descriptive (décrire les caractéristiques du cas) ou explicative (émettre des hypothèses à propos du cas). Lorsque la théorie est explicative, l'étude de cas devrait considérer différentes hypothèses, ce dont doivent tenir compte le choix du devis d'évaluation et le processus de collecte de données (Campbell, 1984, 2009; Yin, 2000b). L'ensemble des éléments mentionnés plus haut caractérise les évaluations par étude de cas, comme le montrera la suite de ce chapitre.

L'étude de cas en tant que méthode complémentaire

Pour juger de la performance d'une intervention, l'évaluation peut se baser sur différentes méthodes, comme celle qui consiste à conduire une étude expérimentale distribuant les sujets entre les groupes contrôle et expérimental de façon aléatoire. Une étude de cas peut alors servir à examiner de plus près un ou plusieurs sujets dans chaque groupe. À titre d'exemple, une évaluation peut porter sur un programme novateur mis en œuvre dans une école. Une étude expérimentale peut répartir les classes entre des groupes contrôle et expérimental pour ensuite comparer les résultats à l'aide d'un même outil, comme le score de rendement des élèves. Une étude de cas a plutôt tendance à se concentrer sur un nombre restreint de classes et à considérer surtout les processus d'enseignement et d'apprentissage. Ainsi, les études de cas permettent de mettre en lumière les processus d'enseignement et d'apprentissage qui caractérisent le programme évalué. Cependant, lorsque les études de cas sont employées comme méthode complémentaire, leurs résultats occupent une place secondaire par rapport à ceux tirés des méthodes principales.

Plusieurs exemples permettent d'illustrer ce fait. Dans le domaine de l'éducation, une évaluation peut porter sur un programme de développement professionnel destiné aux directeurs d'école. Elle peut reposer sur une enquête menée auprès d'un grand nombre de directeurs, alors qu'une étude de cas porte sur un nombre plus restreint. En développement communautaire, l'évaluation d'un programme de développement immobilier

peut faire intervenir une analyse économique des unités de logement et une étude de cas couvrant un plus petit groupe de ménages vivant au sein de ces unités. En santé publique, une évaluation peut prendre en compte les effets d'un nouveau traitement testé dans des cliniques et s'appuyer sur une étude de cas visant à examiner en détail les expériences vécues dans un nombre plus restreint de cliniques.

L'étude de cas en tant que stratégie méthodologique principale

Lorsque l'étude de cas est employée comme stratégie méthodologique principale, le « cas » à l'étude correspond à l'intervention ou au traitement évalué. D'autres données peuvent être collectées par le moyen d'études de cas dont le niveau d'analyse est inférieur (individus ou groupes) ou d'autres types d'études de nature qualitative ou quantitative. On distingue trois types d'évaluation dans lesquels l'étude de cas constitue la démarche principale: 1) l'évaluation des processus; 2) l'évaluation des résultats; 3) l'évaluation des processus et des résultats (voir chapitre 1).

L'évaluation des processus

Comme les études de cas permettent de saisir la complexité d'un cas ainsi que les changements subis par ce dernier au fil du temps², elles sont considérées comme étant tout indiquées pour réaliser des évaluations de processus ou d'implantation. Une étude de cas vise alors à suivre le processus de mise en œuvre et se révèle particulièrement utile si l'intervention présente des caractéristiques d'organisation complexes (voir chapitre 14). En revanche, l'évaluation d'un nouveau médicament ne peut se concentrer sur l'administration unique d'un médicament à un seul patient. Dans un tel cas, l'implantation est dite simple et l'étude de cas n'est pas appropriée. L'étude de cas devrait d'abord déterminer les unités d'analyse principales et secondaires, telles que les individus, les groupes ou les organisations qui chapeautent l'intervention. Ensuite, elle cherche à documenter et à expliquer le processus de mise en œuvre et à fournir des indications sur le potentiel de l'intervention et sur le respect du calendrier. S'il est trop tôt

2. En ayant recours aux entretiens en profondeur et à l'analyse de documents, les études de cas peuvent recueillir des données rétrospectives à propos des événements. Ainsi, le temps de la collecte des données n'est pas toujours celui de l'analyse du cas.

pour pouvoir juger des résultats d'une intervention, l'évaluation de processus ou d'implantation peut constituer la totalité de la démarche évaluative. L'évaluation est alors formative puisque ses résultats permettent d'améliorer ou de réorienter l'intervention. Par exemple, la mise en œuvre d'une intervention complexe peut s'étendre sur plus d'un an et les résultats peuvent se manifester seulement après quelques années. Une étude de cas effectuée durant la première année de la mise en œuvre peut fournir des informations qui seront utiles pour une éventuelle rétroaction.

L'évaluation des résultats

Les études de cas peuvent aussi se concentrer uniquement sur les résultats des interventions. L'évaluation peut avoir pour but de dépouiller l'ensemble des résultats, comme cela est le cas lorsque l'on cherche à étudier les mesures de performance des services publics (Wholey, 1997, p. 131-132). Les études de cas centrées sur les résultats sont également utiles lorsque les principaux résultats sont déjà connus. L'évaluation vise alors à recueillir des données qui rendent compte de l'évolution et de la portée de ces résultats (par exemple, Schwandt, 2009, p. 202). Par exemple, l'évaluation d'une initiative permettant aux élèves de choisir leur école plutôt que d'y être assignés portait sur deux aspects : le rendement des élèves et la capacité de l'intervention à accroître les occasions de s'instruire pour la totalité des parents et des élèves, et non seulement pour une partie d'entre eux (Teske, Schneider, Roch et Marschall, 2000). Afin d'interpréter les résultats, les évaluateurs ont comparé les données provenant des écoles ayant fait l'objet d'une intervention avec celles provenant d'autres écoles du district dans lesquelles il n'y a pas eu d'intervention. Les évaluations par étude de cas peuvent aussi avoir pour but d'expliquer les résultats. Ainsi, une évaluation dans le domaine de l'éducation a montré que les parents ne s'occupaient pas suffisamment des devoirs que leurs enfants avaient à faire à la maison (Yin, 2011, p. 188-192). Ce résultat s'explique, d'abord par le fait que les parents sont trop occupés par leur propre travail et par leurs autres enfants. Toutefois, une analyse plus approfondie des données provenant des études de cas a mis en évidence un autre aspect de l'évaluation : en aidant leurs enfants à obtenir de meilleurs résultats, les parents craignent que la réussite scolaire des enfants ne coupe ceux-ci de leur communauté, caractérisée, depuis plusieurs décennies, par un déclin de l'économie de la population, puisque les enfants auraient ainsi plus de possibilité de quitter leur milieu.

L'évaluation des processus et des résultats

Les études de cas peuvent enfin servir à établir des liens entre les interventions et leurs résultats (Shavelson et Towne, 2002, p. 99-110 ; Mark, 2008, p. 125). Le rôle des études de cas consiste alors à enrichir les évaluations effectuées en leur adjoignant un devis expérimental ou quasi expérimental, même s'il s'agit d'essais randomisés. La principale force des essais randomisés réside dans leur capacité à juger de la relation entre l'intervention et ses résultats et d'attribuer les résultats à l'intervention (par exemple, Bickman et Reich, 2009 ; voir chapitre 6). Cependant, les évaluations par essais randomisés demeurent incomplètes puisqu'elles n'expliquent pas les processus ou les mécanismes permettant aux interventions d'aboutir à des résultats (par exemple, Labin, 2008, p. 101 ; Julnes et Rog, 2009, p. 102-103). Les études de cas viennent combler cette lacune.

À cet égard, les modèles logiques (par exemple, Wholey, 1979 ; Kellogg, 2004 ; voir chapitre 5) jouent un rôle fondamental³. Au départ, le modèle logique peut être reconstitué de façon hypothétique, soit à partir des liens conceptuels qui permettent aux activités (intrants) de produire des effets directs (extrants) et des effets à long terme (impact)⁴. Pour être pertinent, le modèle logique doit opérationnaliser ces liens, c'est-à-dire préciser comment les activités produisent les résultats attendus, ce qui ne peut s'expliquer uniquement par une relation causale. De plus, le modèle logique devrait considérer différentes hypothèses (Yin, 2000b). En effet, « plus l'intervalle entre le début de l'intervention et l'obtention des résultats que l'on souhaite mesurer est long, plus il est difficile de formuler des hypothèses causales différentes » (Julnes et Rog, 2009, p. 110 ; notre traduction).

Malgré le fait qu'il est difficile de réaliser des études de cas permettant d'expliquer comment les interventions produisent des résultats, plusieurs exemples peuvent être cités dans divers domaines : la fermeture d'une base

3. Les modèles logiques sont généralement représentés sous la forme d'un graphique comportant une série de boîtes (désignant un ensemble de variables) reliées par des flèches. La tâche la plus complexe (souvent négligée) est de décrire les flèches qui représentent les mécanismes d'intérêt.

4. Bien que les modèles logiques aient souvent été décrits de façon linéaire, les relations entre les intrants, les extrants et les impacts sont davantage complexes. L'évaluateur peut donc opter pour une approche plus dynamique (par exemple, voir Dyehouse *et al.*, 2009), qu'il est impossible de décrire en détail ici.

militaire d'une petite communauté (Bradshaw, 1999); une initiative visant la revitalisation d'un quartier (Galster, Tatian et Accordino, 2006); des changements dans un système de santé mentale pour enfants (Bickman et Mulvaney, 2006); et une réforme de l'éducation dans un district urbain de grande étendue (Supovitz et Taylor, 2005)⁵.

L'exemple fourni dans la section suivante aidera à mieux comprendre les études de cas axées à la fois sur les processus et les résultats. Il présente la version abrégée d'une évaluation portant sur les activités de prévention de la toxicomanie d'une coalition communautaire (Yin, 2012).

Évaluation des activités de prévention de la toxicomanie d'une coalition communautaire⁶

La coalition était composée d'individus et d'organismes communautaires (centres communautaires, groupes de locataires, associations parentales, etc.). Elle était au service d'une communauté établie dans un quartier où abondaient les magasins de peinture automobile, les centres de recyclage et les logements délabrés et qui était peu à peu devenu un dépotoir de déchets médicaux. Tout cela a favorisé l'émergence de commerces illégaux, tels que la prostitution et le trafic de drogues. À un moment donné, le taux de chômage a été évalué à 47 %. De plus, la concentration de magasins d'alcool était particulièrement élevée, avec 728 commerces détenant un permis d'alcool dans une zone de 40 milles carrés, 10 fois plus que dans les autres quartiers. Contrairement à bon nombre d'organisations communautaires, la coalition était basée sur un modèle d'organisation communautaire qui avait pour but d'amener les citoyens à influencer les prises de décision dans l'arène politique. Elle mettait sur pied des « campagnes d'enjeux ». Une campagne pouvait, à elle seule, couvrir de multiples enjeux tels que le soutien à apporter à des candidats politiques, des projets de loi, des décisions juridiques ou des réglementations.

5. Les auteurs de ces articles n'utilisent pas nécessairement le terme « étude de cas » pour définir leurs évaluations. Toutefois, pour l'ensemble de ces exemples, le principal sujet à l'étude consiste en une seule entité (par exemple, une initiative de revitalisation), les données sont recueillies sur le terrain en fonction du sujet à l'étude et les conclusions portent sur cette seule entité.

6. Les données ont été collectées par le moyen d'entrevues, de revues de publications et de rapports des médias, de séances d'observation et d'un sondage. Ces méthodes ne seront pas décrites dans cette version abrégée.

Les activités de la coalition. Au cours d'une période de troubles, de nombreuses propriétés ont été détruites, dont plus de 200 magasins d'alcool. Un an après les événements, la coalition a organisé une campagne de reconstruction qui avait pour but de faire baisser la consommation d'alcool et de drogues chez les mineurs⁷. L'objectif de la campagne était de reconstruire la communauté et, en même temps, de réduire le nombre de magasins détenant un permis d'alcool. La présence de ces magasins favorisait la consommation d'alcool chez les mineurs, le commerce de drogues ainsi que d'autres activités criminelles. Dans un premier temps, les chefs de la coalition ont rencontré les membres de la mairie, du comité d'urbanisme et du *Alcohol Beverage Control* (ABC) pour discuter du renouvellement des permis d'alcool. La coalition a présenté une pétition de 25 000 signatures demandant un moratoire sur le renouvellement des licences jusqu'à ce que de nouvelles conditions soient établies.

Certaines des nouvelles conditions nécessitaient des changements législatifs. C'est le cas de celles qui exigeaient des peines plus sévères (augmentation de la durée de la suspension du permis) pour les commerçants vendant de l'alcool aux mineurs, ainsi que de celles qui incitaient les commerçants à engager des gardes de sécurité, à améliorer l'éclairage et à surveiller plus étroitement les activités criminelles sur leur territoire. De plus, un nouveau règlement permettrait aux membres du ABC de faire respecter les conditions et d'infliger des sanctions aux commerçants récalcitrants. Au cours de l'année qui a suivi, la coalition a donné à des citoyens la formation nécessaire pour recueillir des preuves, tenir des séances d'information, témoigner devant les instances publiques ou judiciaires et collaborer avec les médias en vue de sensibiliser le public aux problèmes contre lesquels elle luttait. Au bout de quelque temps, le conseil municipal a adopté le nouveau règlement.

7. Cette approche suit un modèle de « santé publique » selon lequel une stratégie de prévention ne réussit parfois que lorsque des modifications sont apportées à l'environnement des individus. Ce modèle contraste avec l'approche individuelle dans laquelle les individus doivent changer leurs comportements, même s'ils sont non pas les responsables des problèmes de santé publique, mais plutôt les victimes. Elle peut s'appliquer dans bon nombre de domaines : élimination des mares d'eau stagnante (pour éviter la propagation des maladies transmises par les moustiques), nettoyage des réservoirs d'eau, suppression des distributeurs de cigarettes dans les endroits fréquentés par les mineurs, cessation de la vente de bières fraîches dans les stations d'essence pour réduire les accidents routiers.

Les commerçants ont d'abord contesté la décision rendue par le conseil municipal et ont demandé à la cour locale d'émettre une injonction, ce qui a été refusé. Les commerçants se sont ensuite tournés vers la cour d'appel qui a également rejeté leur requête afin d'éviter d'aller à l'encontre de la décision rendue par la cour locale. Cela a eu pour effet de renforcer la capacité des administrations municipales à lutter contre les problèmes de criminalité auxquels ils font face. En effet, les villes environnantes se sont prévaluées de cette décision pour imposer des conditions plus strictes à leurs commerçants.

Premiers résultats. L'effet direct des activités de la coalition a d'abord été de permettre aux membres du ABC d'imposer des conditions pour le renouvellement des permis d'alcool. Les membres disposaient également des moyens nécessaires pour assurer le respect des conditions. Il en est résulté par la suite une baisse du nombre de commerces détenant un permis d'alcool. Les nouvelles conditions étant plus strictes, seulement 56 des 200 commerces ont pu renouveler leur permis. Ainsi, non seulement les magasins d'alcool étaient moins nombreux, mais les commerces ayant le permis étaient également plus sécuritaires. L'analyse de la suite des événements, allant de la mise en œuvre de la campagne de reconstruction jusqu'à l'augmentation de l'attention médiatique, a permis de conclure que la coalition a été le principal responsable de la réduction du nombre de commerces détenant un permis d'alcool. L'hypothèse suivant laquelle d'autres acteurs ont également contribué à ces résultats n'a jamais été avancée. Le rôle central de la coalition a été reconnu lorsque d'autres communautés lui ont demandé de les aider à réaliser les mêmes activités.

Résultats quant à la prévention de la toxicomanie. L'étude de cas a évalué les résultats relatifs à la prévention de la toxicomanie en interrogeant un échantillon aléatoire de citoyens et d'élèves du quartier au sujet de leur consommation d'alcool. Les résultats ont ensuite été comparés à des données similaires provenant d'une autre communauté. Cette dernière présentait des caractéristiques économiques et démographiques analogues, mais ne disposait pas d'organismes communautaires ressemblant à la coalition. Les sondages étaient de nature transversale et ont eu lieu avant l'intervention et deux ans après. La comparaison des données provenant de chaque communauté n'a montré aucune différence statistique-

ment significative entre l'évolution des comportements auto-déclarés des adultes ou des jeunes au cours de cette période de deux ans.

Pour montrer d'autres possibilités d'application de la stratégie méthodologique des études de cas dans l'évaluation d'interventions, considérons maintenant le contexte dans lequel s'est inscrit notre recours à des études de cas multiples à deux reprises.

Contexte général des deux exemples d'études de cas en Afrique

Le contexte est celui de l'Afrique de l'Ouest dans les années 1980. Les pays de cette région de l'Afrique ont décidé à cette époque de demander aux patients de payer les soins de santé lorsqu'ils se rendaient dans des établissements sanitaires. Les politiques adoptées dans le domaine de la santé s'inscrivaient dans un plan de réforme du système de santé, lequel fonctionnait plutôt mal, n'était pas utilisé par la population et était sous-financé. De plus, elles visaient à décentraliser le fonctionnement des systèmes de santé en mettant en place des comités locaux de gestion des centres de santé, composés d'habitants des villages environnants, et veillant à ce que le paiement demandé au point de service soit géré localement par ces comités de gestion (COGES). Mais, bien que ces politiques aient rendu possible un approvisionnement constant en médicaments dans les centres de santé primaires, elles ont été incapables d'organiser les systèmes de santé de façon qu'ils répondent aux besoins de la population. Actuellement, ces systèmes de santé sont quasiment délaissés par les populations. L'une des principales raisons est liée au fait que l'on demande aux patients de payer au point de service sans considérer leur capacité de payer et donc sans leur offrir un mode de financement équitable prenant la forme d'une taxe ou d'une assurance universelle (WHO, 2010), ceux qui sont incapables de payer ne vont pas dans les formations sanitaires. Or, les tarifs pratiqués dans ces centres sont sans commune mesure avec les ressources des populations. Ainsi, nous avons calculé que le coût moyen d'une consultation curative (actes et médicaments) pour un enfant de moins de cinq ans dans les centres de santé du nord-est du Burkina Faso était de l'ordre de 1300 F CFA (2,70 \$ CAD). Cette somme représente cinq jours de dépenses (santé, alimentation, école, logement, etc.) pour 46 % de la population du pays considérée comme vivant sous le seuil de pauvreté (INSD, 2009).

Face à cette situation et suivant les recommandations de l'Union africaine, des Nations unies ou de l'OMS, plusieurs pays de l'Afrique de l'Ouest ont imité les pays de l'Afrique australe en établissant des politiques publiques de suppression du paiement des soins au point de service (Meessen *et al.*, 2011 ; Ridde et Morestin, 2011). Les patients ou, du moins, certaines catégories de patients (enfants de moins de cinq ans, femmes enceintes, etc.), n'ont donc plus à payer lorsqu'ils vont dans les centres de santé. Pour compenser le manque à gagner, l'État doit trouver des modalités de remboursement des établissements sanitaires. Pour que ces politiques puissent fonctionner correctement, l'État doit aussi s'assurer que tous les intrants sont disponibles (personnel, médicaments, etc.), que les agents connaissent les modalités de mise en œuvre et que les populations sont au courant. L'objectif est ainsi de rendre les services de santé accessibles à tous et de protéger les ménages contre le risque d'appauvrissement lié aux coûts des soins.

C'est dans ce contexte que deux recherches évaluatives s'appuyant sur la stratégie méthodologique des études de cas multiples ont été conduites. Il s'agit essentiellement ici de montrer comment cette stratégie méthodologique a été employée (Ridde, Diarra et Moha, 2011 ; Ridde et Queuille, 2011).

Étude de cas multiples contrastés sur les processus d'une intervention au Niger

En avril 2006, le gouvernement du Niger a aboli le paiement direct des soins pour les enfants de moins de cinq ans. La formalisation de cette politique figure dans les objectifs du Plan de développement pour la santé (PDS) 2005-2010. Cette politique a trouvé une justification supplémentaire du fait de l'augmentation des taux d'incidence de la malnutrition sévère à la suite de la crise alimentaire de 2005. Il est admis que l'accès aux services de santé aide à combattre la malnutrition.

Le système de santé comporte des districts sanitaires organisés selon une structure pyramidale. Chaque district, administré par une équipe cadre de district (ECD), dispose d'un hôpital employant un médecin et une sage-femme. Le second niveau est composé de centres de santé intégrés (CSI) gérés par des infirmiers. Il y a enfin les postes de santé dans lesquels des agents de santé communautaires (ASC) dispensent des premiers soins payants. Chacun de ces établissements sanitaires recueille les

contributions des utilisateurs. Ces fonds sont conservés et gérés localement par les COGES.

Une « cellule gratuité » coordonne la politique au niveau central. Les établissements sanitaires reçoivent un remboursement pour les services fournis gratuitement, sur la base de taux forfaitaires variant en fonction des types de soins et de structures. Ces fonds permettent aux établissements de santé de s'approvisionner en médicaments dans les pharmacies populaires, qui sont des satellites de la centrale nationale d'achat (ONPPC).

En 2009, soit deux ans après la mise en place de la politique à l'échelle nationale, une évaluation a été menée en vue de rendre compte de la mise en œuvre de la politique dans deux districts.

L'approche méthodologique adoptée est celle des études de cas multiples avec plusieurs niveaux d'analyse imbriqués (Yin, 2009). Les cas sont les districts de santé. Ayant choisi d'adopter une démarche heuristique et comparative, nous avons sélectionné deux cas contrastés, puisque « des cas multiples sont considérés comme plus robustes », comme le précise Yin (2009). Le contraste s'observe entre, d'une part, le soutien d'une ONG internationale dont a bénéficié l'un des deux districts et qui est susceptible d'avoir facilité la mise en œuvre de la politique, et, d'autre part, des différences qui tiennent au contexte. Seul le district de Keïta a été appuyé par une ONG depuis 2006. Cet appui ne s'est pas limité aux aspects organisationnels de la politique d'abolition. L'ONG a en outre fait de la gestion, veillé à la qualité des soins et à l'évacuation des malades, formé du personnel, rénové des centres de santé, etc. Pour des raisons financières et logistiques, nous avons décidé de procéder à une comparaison avec un des districts situés à proximité de Keïta, soit celui d'Abalak. Ce district était celui qui offrait le meilleur contraste sur le plan géographique (population et climat).

L'étude a utilisé des données qualitatives collectées entre janvier et mars 2009. Des analyses documentaires et des enquêtes de terrain de type socio-anthropologique (Olivier de Sardan, 2008) ont été réalisées dans tous les établissements sanitaires des deux districts ($n = 23$), y compris dans les hôpitaux de district. En vue d'obtenir une triangulation de données, nous avons mené des entrevues individuelles approfondies et organisé des *focus groups* avec des informateurs clés rangés en trois catégories (Monnier et Spennlehauser, 1992) : 1) les *actants*, c'est-à-dire les personnes qui ont participé à la mise en œuvre de la politique (agents de santé, membres du personnel de l'ONG, membres des COGES, etc.) ; 2) les *réac-*

tants, c'est-à-dire les bénéficiaires de l'action ; 3) les *légitimants*, soit les décideurs concernés. Au total, nous avons rencontré 221 personnes, 40 % d'entre elles étant des femmes.

Toutes les entrevues individuelles et de groupe ont été menées par les chercheurs à l'aide du même canevas. Les entrevues conduites dans la langue locale ont été réalisées par un des chercheurs, puis traduites en français de manière que tous puissent participer à leur analyse. Dans les centres de santé et les équipes de district, tous les agents de santé et les décideurs présents et disponibles ont été interrogés. Quant aux bénéficiaires, nous les avons sélectionnés au hasard parmi les personnes présentes à chacune de nos visites. Dans certains villages, des associations de femmes nous ont aidés à organiser des *focus groups* avec des femmes qui ne se trouvaient pas dans le centre de santé. Pour répondre aux exigences de la triangulation, nous avons veillé à ce que les profils des personnes sélectionnées soient diversifiés. Sur chaque site, nous avons mis fin à l'étape des entrevues lorsque nous estimions avoir atteint un point de saturation empirique. Les entrevues ont fait l'objet d'une prise de notes systématique, et un grand nombre d'entre elles ($n = 167$) ont été enregistrées et transcrites. Les données ont ensuite été analysées suivant l'approche de l'analyse de contenu.

Les résultats, que nous résumons ici brièvement (voir Ridde *et al.*, 2011), indiquent que tout le monde reconnaît que la décision d'abolir le paiement direct des soins aux enfants de moins de cinq ans au Niger est à l'origine largement politique. Bien que cette décision soit techniquement pertinente pour lever l'obstacle pécuniaire et améliorer l'accès aux soins des enfants, elle n'est peut-être pas adaptée au contexte. Dans le cas du Niger, il y a eu un manque de préparation technique, et le financement s'est révélé insuffisant, tout cela sur fond d'instabilité politique. La présente étude de cas montre que la garantie d'un bon accès aux médicaments et le renforcement du système d'approvisionnement sont des éléments indispensables au succès d'une telle politique. Il apparaît en outre qu'on n'a pas vraiment tiré un enseignement d'un projet pilote mené par une autre ONG dans deux districts (Ridde et Diarra, 2009) avant que la politique ne soit étendue à tout le pays.

En ce qui concerne les limites méthodologiques, il convient de préciser que, puisque cette étude portait sur deux districts (deux cas) seulement du Niger, il était impossible d'étendre nos conclusions à l'ensemble du

territoire de ce pays. En réalité, les deux districts n'ont pas été sélectionnés pour leur représentativité de l'ensemble du Niger – une manière de procéder qui dépassait nos moyens –, mais plutôt parce qu'ils offraient deux situations suffisamment contrastées pour nous permettre d'arriver à une bonne compréhension du phénomène étudié. De plus, nos conclusions ont acquis une fiabilité supplémentaire du fait de la profondeur de notre analyse, de la triangulation des données et de la validation des résultats par les principaux acteurs concernés. L'ONG présente à Keïta n'est intervenue d'aucune manière dans le processus d'évaluation. En outre, les résultats de cette étude s'accordent en grande partie avec les analyses réalisées par les autorités politiques (Tari Bako, 2010) et par les experts étrangers (Louis, Nachtigal, Kourguéni et Issa, 2008), de même qu'avec les résultats d'études en cours (Olivier de Sardan et Ridde, 2011; Ousseni, 2010). Le recours à deux cas contrastés permet de voir si des réplifications logiques résistent à la confrontation avec les données empiriques. Cela renforce par ailleurs la « généralisation analytique » (Yin, 2009) de la difficulté liée à la mise en œuvre de la politique d'abolition.

Études de cas multiples avec plusieurs niveaux d'analyse imbriqués concernant la mise en œuvre et les résultats d'interventions dans six pays de l'Afrique de l'Ouest

En attendant que les études scientifiques sur ces politiques soient publiées, il a été décidé d'aider les responsables de leur mise en œuvre dans un exercice de capitalisation de leur expérience⁸. En effet, les connaissances empiriques dont les « *implementors* » (Pressman et Wildavsky, 1984) disposent peuvent aussi être utilisées en vue de mieux comprendre ces politiques sans attendre que les scientifiques, dont le travail est souvent long, produisent et rendent publiques des données probantes. Aussi, des équipes de personnes responsables de la mise en place de ces politiques dans six pays d'Afrique de l'Ouest (Bénin, Burkina Faso, Mali, Niger, Togo, Sénégal) ont-elles été regroupées afin de réaliser une capitalisation nationale (étude de cas unique), puis transversale (étude de cas multiples), concernant ces politiques. Ces équipes se sont réunies plusieurs fois et ont travaillé pendant

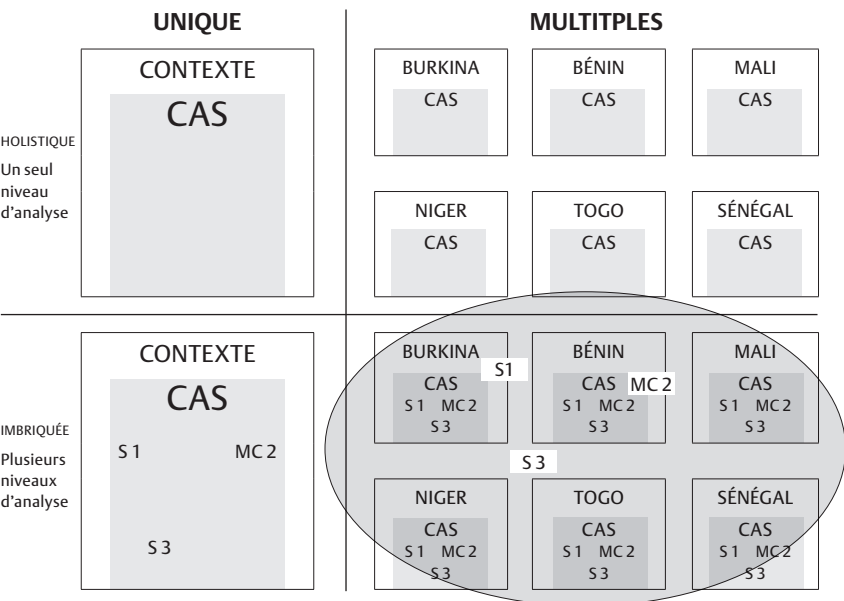
8. Merci à Ludovic Queuille et Yamba Kafando qui ont été les principaux organisateurs de ce projet financé par l'Union européenne (ECHO) en collaboration avec l'ONG allemande HELP.

environ un an (2010-2011) sur une étude de cas unique pour chaque pays⁹, le cas étant la politique d'exemption du paiement des soins appliquée dans chacun de ces pays. Ainsi, le Sénégal a étudié la politique d'exemption du paiement des soins pour les personnes âgées de plus de 60 ans (seule politique du genre en Afrique), tandis qu'au Togo on a considéré la politique d'exemption du paiement des traitements antirétroviraux contre le VIH/sida. Les données utilisées pour ces études de cas unique étaient variées : données secondaires prises dans des rapports d'experts, de chercheurs ou de l'administration publique, données primaires obtenues au moyen d'un questionnaire ou d'entrevues approfondies, etc.

L'analyse transversale effectuée par la suite provient donc de l'analyse de ces études de cas unique. Ainsi, parmi les quatre types d'études de cas possibles (figure 10.1), la stratégie méthodologique retenue dans ce processus de capitalisation inter-pays est celle de l'étude de cas multiples à plusieurs niveaux d'analyse imbriqués (cases encadrées de bleu) (Yin, 2009).

FIGURE 10.1

Les différentes formes d'études de cas (en encadré, le choix pour l'analyse transversale)



9. Voir Kratochwill et ses collaborateurs (2010) pour une autre façon de présenter la démarche des études de cas unique.

Les cas sont constitués des politiques d'exemption du paiement des soins mises en œuvre dans six pays. Le choix des cas dans chaque pays a été en quelque sorte instrumental, il a été effectué en fonction de l'utilité que les membres des équipes pays pouvaient tirer des résultats de l'étude (voir Patton et LaBossière, chapitre 8).

L'analyse transversale, donc l'étude de cas multiples, rend compte de la synthèse des connaissances acquises dans les pays suivant plusieurs niveaux d'analyse imbriqués. Les différents niveaux d'analyse, dont quelques exemples sont présentés à la figure 10.1 (S1, S3, MC2; voir le tableau 10.1 pour les dénominations), sont empruntés au cadre analytique de l'étude qui s'inspire de l'approche d'évaluation *realist* (Pawson et Tilley, 1997; Ridde *et al.*, 2012) décrite au chapitre 14. Les niveaux d'analyse correspondent, d'une part, aux stratégies suivies au moment de l'application de ces politiques, et, d'autre part, aux attitudes et aux réactions de la population face à ces politiques. Autrement dit, il s'agissait de documenter les activités organisées et les mécanismes cognitifs (Astbury et Leeuw, 2010) mobilisés par les acteurs sociaux concernés par ces politiques. Les éléments constituant ces deux catégories de niveaux d'analyse ont été déterminés à la suite d'une revue des écrits traitant de ces politiques publiques et des interventions des organisations non gouvernementales (Robert, 2011). Ainsi, les responsables de l'application des politiques dans chacun des pays concernés ont été en mesure de collecter des données nationales et de classer les résultats de leurs expériences en fonction de 10 éléments, compris comme des niveaux d'analyse, de ces deux catégories (tableau 10.1).

Ainsi, cette étude de cas multiples à plusieurs niveaux d'analyse imbriqués s'appuyait sur trois sources de données.

TABLEAU 10.1

**Les niveaux d'analyse employés
dans l'étude de cas multiples**

Stratégies de mise en œuvre	Attitudes des individus
gestion (S1)	motivation et satisfaction du personnel (MC1)
communication (S2)	relation soignant/utilisateur (MC2)
suivi/évaluation/coordination (S3)	satisfaction des bénéficiaires (MC3)
participation communautaire (S4)	perception des effets par les bénéficiaires (MC4)
prise en charge et orientation du patient (S5)	adaptation du personnel fournissant des soins de santé (MC5)

La première source est constituée par les discussions et les interactions permanentes entre les auteurs de l'analyse transversale et les membres des équipes des pays concernant leur étude de cas unique. Ces discussions ont eu lieu dans les trois ateliers qui réunissaient tous les protagonistes du projet. Au cours de ces rencontres collectives, les auteurs des études de cas unique ont présenté oralement leurs travaux, puis il y a eu des discussions de groupe, et d'autres travaux de groupe ont permis aux membres des équipes des différents pays de poursuivre la rédaction de leur étude. Le contenu de ces discussions a été pris en compte dans l'analyse transversale.

La deuxième source est constituée par les versions intermédiaires et finales des études de cas. Ainsi, les études de cas des pays ont toutes été lues en entier plusieurs fois, ce qui était nécessaire pour bien comprendre et analyser leur contenu.

La troisième source de données est constituée par l'analyse transversale amorcée par les auteurs des études de cas unique par pays. Dans l'atelier final du projet (mai 2011), les membres des équipes des différents pays ont formé plusieurs groupes de travail afin d'effectuer un début d'analyse transversale par groupe de trois pays. La présentation de ces travaux en séance plénière ainsi que la version préliminaire de l'analyse transversale fondée sur les deux premières sources de données ont aussi permis de poursuivre la réflexion analytique. Toutes ces discussions ont fait l'objet de prises de notes systématiques et ont été enregistrées en vue d'une analyse subséquente.

L'une des limites méthodologiques – mais, à un autre point de vue, un avantage – des processus de capitalisation est que les nombreuses données endogènes utilisées proviennent de propos énoncés par les acteurs, qui sont souvent aussi les premiers responsables de la mise en œuvre des politiques étudiées. Les conflits d'intérêts, surtout dans des contextes africains où il est relativement difficile de mener une démarche réflexive et critique (Ndiaye, 2009), sont évidemment possibles. Cependant, d'une part, les auteurs des études de cas ne s'en sont pas toujours tenus à un discours lénifiant et, d'autre part, les auteurs de l'analyse transversale, qui connaissaient parfaitement les conditions dans lesquelles étaient appliquées ces politiques, ont été capables de rédiger l'étude de cas multiples avec le recul critique nécessaire.

L'analyse des résultats de cette étude de cas multiples montre que, dans l'ensemble, les processus de mise en œuvre de ces politiques d'exemption

sont plutôt laborieux, par-delà les effets bénéfiques qui ont été décrits dans d'autres publications (Ridde et Morestin, 2011) et perçus ici (MC3). Ces difficultés dans la mise en application ont aussi été largement signalées dans d'autres pays d'Afrique (Meessen *et al.*, 2011). Ce n'est donc pas nouveau. Mais les études de cas effectuées dans ces six pays d'Afrique de l'Ouest confirment l'importance, une fois les décisions prises, d'accorder une attention particulière aux processus de mise en œuvre afin que les politiques aient tous les effets positifs escomptés. Mieux, l'approche séquentielle et linéaire des politiques publiques suivant laquelle l'émergence est suivie de la formulation et de la mise en œuvre doit être repensée de façon à obtenir une vision plus consensuelle, tourbillonnaire ou interactive (Grindle et Thomas, 1991; Monnier, 1992; Ridde, 2010). Car si le principe de l'exemption du paiement paraît partout apprécié, les difficultés auxquelles se heurte sa mise en œuvre pourraient aboutir à une remise en cause de ses fondements.

Conclusion

Dans ce chapitre, nous avons décrit une des utilisations possibles des études de cas dans le domaine de l'évaluation de programmes. Encore une fois, il ne s'agit qu'une des multiples approches possibles, mais nous avons voulu, dans ce texte, présenter quelques concepts clés ainsi que des exemples de leur application dans des contextes très différents. Il reste maintenant à employer cette stratégie méthodologique et à suggérer à ceux ou celles qui l'utilisent de réfléchir aux moyens de la développer.

Références

- ASTBURY, B. et F. L. LEEUW (2010), « Unpacking black boxes: Mechanisms and theory building in evaluation », *American Journal of Evaluation*, 31(3), 363-381.
- BICKMAN, L. et S. MULVANEY (2005), « Large scale evaluation of children's mental health services: The Ft. Bragg and Stark County Studies », dans STEELE, R. et M. ROBERTS (dir.), *Handbook of Mental Health Services for Children, Adolescents, and Families*, New York, Kluwer Academic/Plenum.
- BICKMAN, L. et S. M. REICH (2009), « Randomized controlled trials: A gold standard with feet of clay? », dans DONALDSON, S. I., C. A. CHRISTIE et M. M. MARK (dir.), *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* (p. 51-77), Thousand Oaks, Sage.

- BRADSHAW, T. K. (1999), « Communities not fazed : Why military base closures may not be catastrophic », *Journal of American Planning Association*, 65, 193-206.
- CAMPBELL, D. T. (2009), « Foreword », dans Yin, R. K., *Case Study Research : Design and Methods*, 4^e éd., Thousand Oaks, Sage.
- CRONBACH, L. J. *et al.* (1980), *Toward Reform in Program Evaluation*, San Francisco, Jossey-Bass.
- DATTA, L. E. (1997), « Multimethod evaluations : Using case studies together with other methods », dans CHELIMSKY, E. et W. R. SHADISH (dir.), *Evaluation for the 21st Century : A Handbook* (p. 344-359), Thousand Oaks, Sage.
- DYEHUSE, M. *et al.* (2009), « A comparison of linear and systems thinking approaches for program evaluation illustrated using the Indiana Interdisciplinary GK-12 », *Evaluation and Program Planning*, 32, 187-196.
- GALSTER, G., P. TATIANET et J. ACCORDINO (2006), « Targeting investments for neighborhood revitalization », *Journal of the American Planning Association*, 72, 457-474.
- GRINDLE, M. S. et J. W. THOMAS (1991), *Public Choices and Policy Change. The Political Economy of Reform in Developing Countries*, Baltimore et Londres, The Johns Hopkins University Press.
- INSD (2009), *Annuaire statistique. Édition 2008*, Ouagadougou, INSD.
- JULNES, G. et D. ROG (2009), « Evaluation methods for producing actionable evidence : Contextual influences on adequacy and appropriateness of method choice », dans DONALDSON, S. I., C. A. CHRISTIE et M. M. MARK (dir.), *What Counts as Credible Evidence in Applied Research and Evaluation Practice ?* (p. 96-131), Thousand Oaks, Sage.
- KELLOGG FOUNDATION (2004), *Using Logic Models to Bring Together Planning, Evaluation, and Action : Logic model Development Guide*, Battle Creek, à compte d'auteur.
- KRATOCHWILL, T. R. *et al.* (2010), « Single-case designs technical documentation », What Works Clearinghouse website : http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf
- LABIN, S. N. (2008), « Research syntheses : Toward broad-based evidence », dans SMITH, N. L. et P. R. BRANDON (dir.), *Fundamental Issues in Evaluation* (p. 89-110), New York, Guilford.
- LOUIS, J.-P., J. NACHTIGAL, I. A. KOURGUÉNI et M. S. ISSA (2008), *Évaluation externe à mi-parcours du PDS 2005-2010* (p. 133), Programme FED.
- MARK, M. M. (2008), « Building a better evidence base for evaluation theory », dans SMITH, N. L. et P. R. BRANDON (dir.), *Fundamental Issues in Evaluation* (p. 111-134), New York, Guilford.
- MEESSEN, B. *et al.* (2011), « Removing user fees in the health sector : A review of policy processes in six sub-Saharan African countries », *Health Policy and Planning* (Supplement 2), 26, ii16-ii29.

- MONNIER, E. (1992), *Évaluations de l'action des pouvoirs publics. Du projet au bilan*, Paris, Économica.
- MONNIER, E. et V. SPENLEHAUER (1992), « L'évaluation dans le triangle de la décision. Opinions singulières et processus pluraliste », *Politiques et management public*, 10(3), 61-82.
- NDIAYE, A. (dir.) (2009), *Chercheurs et décideurs d'Afrique. Quelles synergies pour le développement ?*, Dakar, CODESRIA/CRDI.
- OLIVIER DE SARDAN, J.-P. (2008), *La rigueur du qualitatif. Les contraintes empiriques de l'interprétation socio-anthropologique*, Louvain-la-Neuve, Academia Bruylant.
- OLIVIER DE SARDAN, J.-P. et V. RIDDE (2011), *Une comparaison provisoire des politiques d'exemption de paiement dans trois pays sahéliens (Burkina Faso, Mali, Niger)* (p. 32), Niamey, LASDEL/CRCHUM.
- OUSSENI, A. (2010), *Une politique publique de santé au Niger. La mise en place d'exemptions de paiement des soins en faveur des femmes et des enfants* (p. 58), Niamey, LASDEL.
- PAWSON, R. et N. TILLEY (1997), *Realistic Evaluation*, Londres, Thousand Oaks, Sage.
- PRESSMAN, J. L. et A. WILDAVSKY (1984), *Implementation. How Great Expectations in Washington Are Dashed in Oakland* (3^e éd.), Berkeley, Los Angeles, Londres, University of California Press.
- RIDDE, V. (2010), « Multiple streams theory », dans T. Kurian (dir.), *The Encyclopedia of Political Science*, CQ Press.
- RIDDE, V. et A. DIARRA (2009), « A process evaluation of user fees abolition for pregnant women and children under five years in two districts in Niger (West Africa) », *BMC Health Services Research*, 9(89).
- RIDDE, V. et F. MORESTIN (2011), « A scoping review of the literature on the abolition of user fees in healthcare services in Africa », *Health Policy and Planning*, 26(1), 1-11.
- RIDDE, V. et L. QUEUILLE (dir.) (2012), *Capitalisation des politiques publiques d'exemption du paiement dans six pays d'Afrique de l'Ouest*, Montréal, CRCHUM.
- RIDDE, V., A. DIARRA et M. MOHA (2011), « User fees abolition policy in Niger : Comparing the under five years exemption implementation in two districts », *Health Policy*, 99, 219-225.
- RIDDE, V. et al. (2012), « L'approche Realist à l'épreuve du réel », *Revue canadienne d'évaluation de programmes*, numéro spécial sur l'évaluation des interventions complexes, (26) 3.
- ROBERT, É. (2011), « État des lieux des connaissances disponibles sur les politiques et les programmes d'exemption des paiements directs des soins en Afrique de l'Ouest : Bénin, Burkina Faso, Ghana, Guinée, Mali, Niger et Sénégal »,

- Montréal, CRCHUM/ONG HELP/ECHO. Disponible à : www.medsp.umontreal.ca/vesa-tc/ressrc.htm
- SCHWANDT, T. A. (2009), « Toward a practical theory of evidence for evaluation », dans DONALDSON, S. I., C.A. CHRISTIE et M. M. MARK (dir.), *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* (p. 197-212), Thousand Oaks, Sage.
- SHAVELSON, R. et L. TOWNE (2002), *Scientific Research in Education*, Washington, D. C., National Academy Press.
- SUPOVITZ, J. A. et B. S. TAYLOR (2005), « Systematic education evaluation: Evaluating the impact of system wide reform in education », *American Journal of Evaluation*, 26, 204-230.
- TARI BAKO, I. F. (2010), *Présentations thématiques. Chirurgie de districts et gratuité des soins*, Niamey, Direction de l'organisation des soins, janvier.
- TESKE, P., M. SCHNEIDER, C. ROCHET et M. MARSCHALL (2000), « Public school choice: A status report », dans RAVITCH, D. et J. P. VITERITTI (dir.), *Lessons from New York City Schools* (p. 313-338), Baltimore, Johns Hopkins University Press.
- WHO (2010), *The World Health Report: Health Systems Financing: The Path to Universal Coverage* (p. 106), Genève, WHO.
- WHOLEY, J. S. (1979), *Evaluation: Performance and Promise*, Washington, D. C., The Urban Institute.
- WHOLEY, J. S. (1997), « Trends in performance measurement: Challenges for evaluators », dans CHELIMSKY, E. et W. R. SHADISH (dir.), *Evaluation for the 21st Century: A Handbook* (p. 124-133), Thousand Oaks, Sage.
- YIN, R. K. (2009), *Case Study Research: Design and Methods*. (4^e éd.), Los Angeles, Sage Publications.
- YIN, R. K. (2000a), « Case study evaluations: A decade of progress? », dans STUFFLEBEAM, D. L., G. F. MADAUS et T. KELLAGHAN (dir.), *Evaluation Models: Viewpoints on Educational and Human Services Evaluation* (2^e éd.) (p. 185-193), Boston, Kluwer.
- YIN, R. K. (2000b), « Rival explanations as an alternative to “reforms as experiments” », dans BICKMAN, L. (dir.), *Validity and Social Experimentation: Donald Campbell's Legacy* (p. 239-266), Thousand Oaks, Sage.
- YIN, R. K. (2009), *Case Study Research: Design and Methods* (4^e éd.), Los Angeles, Sage.
- YIN, R. K. (2011), *Qualitative Research from Start to Finish*, New York, Guilford Press.
- YIN, R. K. (2012), *Applications of Case Study Research* (3^e éd.), Los Angeles, Sage.

11

Une introduction à l'évaluation économique des programmes

Sophie Witter et Nadia Cunden

Ce chapitre donne un aperçu des différents types d'évaluation économique utilisés pour éclairer les décisions politiques qui doivent prendre en compte l'efficacité des programmes (chapitre 1). Il traite de l'analyse coût-efficacité, notamment dans le domaine de la santé, car ce genre d'analyse est des plus simples et est très utilisé.

Le chapitre est divisé en cinq sections. La section 1 définit ce qu'est l'économie de la santé et montre l'utilité de l'évaluation économique dans le domaine. La section 2 décrit les différentes méthodes d'évaluation économique. La section 3 décrit les étapes à suivre dans une analyse coût-efficacité. La section 4 traite des questions méthodologiques liées à l'application de l'analyse coût-efficacité. La section 5 présente la conclusion.

L'économie de la santé

L'économie a rapport à la production, à la gestion, à la consommation et à la répartition des biens au sein de la société. Dans un marché parfait, les ressources sont allouées de manière à optimiser le bien-être social. Cela n'est possible que si certaines conditions sont remplies. Il faut notamment :

- qu'il y ait atomicité du marché, c'est-à-dire qu'il y ait des acheteurs et des vendeurs en grand nombre et qu'ils soient de petite taille (or il y a

un grand nombre d'acheteurs et de vendeurs sur le marché, et aucun ne peut fixer les prix);

- qu'il y ait libre entrée et libre sortie; les acheteurs et les vendeurs doivent pouvoir entrer et sortir librement du marché (donc il doit y avoir absence de barrières réglementaires, tarifaires ou de coûts importants);
- qu'il y ait homogénéité des produits (les produits circulant sur le marché sont de nature et de qualité comparables);
- qu'il y ait transparence de l'information (quand l'information est parfaite, les prix et la qualité des produits sont connus de tous les acheteurs et vendeurs sur le marché).

En se basant sur ces exigences, il est clair que le marché de la santé est imparfait, car l'offre et la demande dans ce domaine n'obéissent pas aux règles. *Le numerus clausus*, règle qui fixe un nombre limité de personnes, limite le nombre de médecins et les réglementations limitent l'entrée sur le marché des soins. Les soins et les tarifs de santé ne sont pas homogènes et il y a disproportion entre les informations dont dispose le patient et celles que possède le médecin, car ce dernier, par sa formation, connaît mieux l'état de santé du patient (Stiglitz *et al.*, 2007). Il est donc nécessaire que les gouvernements interviennent, s'occupent du financement et de la régulation du marché de la santé ainsi que de la provision des prestations de santé. Les services de soins de santé sont également caractérisés par des externalités positives (par exemple, lorsqu'un vaccin protège les personnes non vaccinées entrant en contact avec une personne qui vient d'être vaccinée) et négatives (par exemple, lorsque l'inobservance d'un traitement antituberculeux provoque la transmission de la tuberculose aux membres de la famille ou de la communauté). Tous ces facteurs influent sur les décisions concernant la manière d'estimer le coût d'un service et la quantité de services à fournir.

Dans l'ensemble, l'économie de la santé fournit une aide aux décideurs politiques quant à la répartition des ressources limitées imparties au système de santé et aux soins de santé. Elle fournit des outils qui permettent d'éclairer les décisions relatives aux types de services à donner, aux catégories de personnes qui peuvent en bénéficier et à la manière de les délivrer. L'évaluation économique nécessite une comparaison systématique des différents types de solutions en fonction de leurs coûts et de leurs effets.

L'évaluation économique consiste à cerner, à évaluer et à comparer les coûts et les conséquences des différentes solutions possibles (Drummond

et al., 1997). Les méthodes d'évaluation économique permettent de comparer entre eux les différents programmes de santé en vue de déterminer lequel justifie un investissement. Les évaluateurs considèrent donc les coûts d'opportunités d'une intervention, c'est-à-dire ce qui aurait pu être obtenu avec les mêmes ressources si elles n'avaient pas été investies dans cette intervention.

L'évaluation économique, cependant, ne permet pas de donner des réponses à toutes les questions. Elle peut être utile pour évaluer les intrants et les extrants dans différents services, et pour les comparer de manière objective les uns avec les autres. Mais quand il s'agit de prendre des décisions, il est nécessaire de tenir compte des priorités sociales ainsi que de la disponibilité et de la régularité du financement. Les économistes de la santé doivent également travailler en étroite collaboration avec les épidémiologistes, les responsables de la santé publique et les chercheurs qui mènent des essais cliniques dont les résultats peuvent amener à concevoir des programmes plus efficaces et plus universels pour réviser les calculs économiques (Witter *et al.*, 2000).

Les différents types d'évaluation économique

Les principaux types d'évaluation économique sont l'analyse de minimisation des coûts, l'analyse coût-efficacité, l'analyse coût-utilité et l'analyse coûts-bénéfice. Ils se différencient principalement en ce qui concerne la mesure d'effet (tableau 11.1).

TABLEAU 11.1

Principales méthodes d'évaluation économique

Méthode	Estimation des résultats	Conditions d'application
Minimisation des coûts	Types de résultat identique	N'est possible que si les interventions ont des effets identiques.
Coût-efficacité	Unités naturelles (cas guéris, années de vie gagnées)	Ne peut être utilisée que pour comparer les interventions avec des résultats semblables.
Coût-utilité	Unités naturelles amalgamées (années de vie ajustées pour la qualité)	Permet les comparaisons entre un large éventail de résultats et peut combiner les effets de morbidité et de mortalité.
Coût-bénéfice	Unités monétaires	Les résultats peuvent être comparés avec d'autres interventions même dans d'autres secteurs que la santé.

L'analyse de minimisation des coûts

L'analyse s'intéresse aux seuls coûts. Elle est utilisée lorsque les résultats des différentes interventions sont les mêmes et que le but de l'évaluation est de trouver la solution la moins onéreuse. Comme dans les autres formes d'évaluation économique, les interventions sont évaluées selon le prix du marché. Autant que possible, on prend en compte les facteurs sociaux en évaluant l'impact, non seulement sur les services de soins de santé, mais également sur les patients, leurs familles et la société (Robinson, 1993).

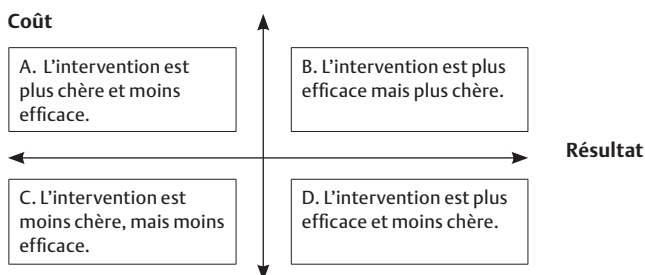
L'analyse coût-efficacité

L'analyse coût-efficacité considère à la fois les coûts et les effets des interventions sur la santé, elle compare les différentes interventions possibles entre elles ou avec les interventions actuelles. La comparaison s'appuie en principe sur un ratio coût-efficacité incrémentiel. Cela permet d'évaluer les coûts supplémentaires liés à l'introduction ou au renforcement d'un service, comparés à ses avantages supplémentaires attendus.

Une analyse coût-efficacité donne généralement une estimation du coût par unité d'effet ou des effets par unité de coût (Drummond *et al.*, 1997), tel que le coût par cas de maladie détecté ou le coût résultant de la mise en place d'un traitement qui réduit le nombre de jours de maladie (Hull *et al.*, 1981; Sculpher et Buxton, 1993). Le graphique de coût-efficacité (figure 11.1) illustre une façon de présenter aux décideurs les ratios de coût-efficacité

FIGURE 11.1

Plan coût-efficacité



Source : Adapté de Fox-Rushby et Cairns (2005).

liés aux diverses solutions possibles. L'intervention tombant dans le quadrant A serait clairement rejetée et celle qui est dans le quadrant D serait adoptée dans la plupart des cas. L'intervention en C pourrait être adoptée si les budgets étaient limités et l'intervention en B pourrait être choisie si le problème était crucial et si les ressources financières le permettaient.

L'analyse coût-efficacité est le type d'évaluation économique le plus utilisé dans le domaine de la santé. Cependant, celle-ci convient uniquement quand les résultats peuvent être comparés au moyen d'une seule mesure.

L'analyse coût-bénéfice

L'analyse coût-bénéfice mesure les coûts et les effets de deux solutions sous le rapport financier. L'analyse coût-bénéfice est souvent celle que favorise un ministère des Finances, car elle permet une comparaison de la valeur sociale à travers les différents secteurs. Elle tient compte également d'un jugement absolu afin d'établir si les fonds investis sont supérieurs aux bénéfices. Aucune autre méthode d'évaluation économique ne tient compte de cette décision simple et de cette règle absolue (l'investissement est valable si les bénéfices nets actuels¹ sont supérieurs à zéro). Cependant, il est difficile de chiffrer les bénéfices liés à la santé. Une approche parfois utilisée consiste à évaluer si la population est prête à payer un service donné. Toutefois, du fait de la nature hypothétique de ces questions, les résultats apparaissent discutables. De plus, l'évaluation tiendra probablement compte des revenus d'un individu ou de sa « productivité économique » (Hurley, 2000). Cela peut parfois favoriser les individus plus aisés.

L'analyse coût-utilité

L'analyse coût-utilité considère la préférence que les personnes attachent aux états de santé. Les préférences sont généralement mesurées sur une échelle de 0 à 1, 0 correspondant au décès et 1 à un état de parfaite santé.

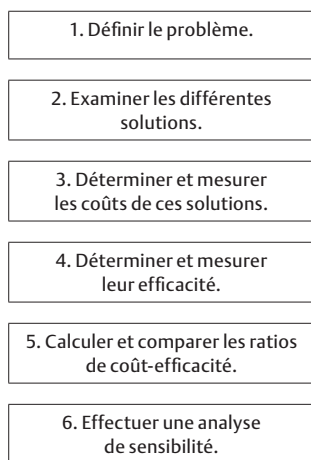
L'analyse coût-utilité utilise généralement les années de vie gagnées ajustées pour la qualité (QALYs), les années de vie en bonne santé ou les

1. Les bénéfices nets actuels sont les bénéfices futurs d'une intervention qui ont été convenablement actualisés pour tenir compte de la préférence de temps, et dont les coûts futurs ont été correctement actualisés et soustraits.

années de vie corrigées de l'incapacité (AVCI) comme mesures des résultats. Cette méthode d'évaluation est utilisée quand une intervention agit sur la mortalité et la morbidité. Par exemple, les traitements contre le cancer augmentent le nombre d'années de vie, mais, dans de nombreux cas, ils réduisent la qualité de la vie. Cependant, certaines suppositions liées à l'analyse coût-utilité sont controversées. Par exemple, il est nécessaire de supposer, dans ce type d'analyse, qu'une année de vie gagnée ajustée pour sa qualité a la même valeur pour tout le monde, mais cela entraîne une simplification. L'analyse coût-utilité s'inscrit dans une optique utilitaire. Par exemple, il peut s'agir d'améliorer de 10 % l'état de santé d'une personne pendant 10 ans, ce qui équivaut à donner une année supplémentaire de vie à 100 %. Ce « modèle mathématique » peut ne pas correspondre aux valeurs et aux perceptions réelles de la population. En ce qui concerne les AVCI, les pondérations normatives données à l'espérance de vie et à l'incapacité ont été critiquées, et certains auteurs ont fait valoir que les jugements de valeur formulés proviennent principalement des populations instruites (Morris *et al.*, 2007).

Les étapes d'une analyse coût-efficacité

Le schéma suivant indique les étapes à suivre dans la réalisation de tous les types d'analyse coût-efficacité :



Source : Adapté de Reynolds et Gaspari (1985).

Les coûts des services de santé comprennent notamment des coûts d'investissement (comme l'achat de véhicules, dont les coûts peuvent être répartis en fonction de leur durée de vie utile). De plus, il peut aussi y avoir des coûts de démarrage non récurrents et limités dans le temps, comme les coûts liés à la formation (voir l'étude de cas 1). Il y a aussi presque toujours des coûts récurrents comme ceux qui sont liés au carburant, aux salaires et aux communications. Si ces coûts sont propres au programme, ils sont relativement simples à calculer. Cependant, s'ils sont partagés, ils doivent être distribués selon une clé de répartition entre les interventions en question à partir d'un indicateur approprié (tel que le temps moyen mis à faire une intervention). Il peut aussi y avoir de nouveaux coûts de gestion attachés à une nouvelle intervention ou à un élargissement du programme. D'autre part, les économies sur les coûts qui ont été réalisées (par exemple, les traitements futurs qui sont évités) peuvent être soustraites pour obtenir les coûts nets de l'intervention.

Dans le cas où on prend en compte l'aspect social, les coûts pour les patients sont également considérés – (par exemple, les frais de transport et les pertes de revenu entraînées par une intervention médicale. Ces coûts doivent d'abord être estimés, puis calculés, en utilisant, par exemple, les sommes quotidiennes reçues en salaire d'après le marché local (si on ne connaît pas le revenu individuel). Le choix du coût (pour le patient ou pour le prestataire) dépend de l'optique dans laquelle s'inscrit l'analyse. Si l'évaluation a pour but d'informer les décideurs publics, le point de vue des prestataires peut être suffisant. Mais en général, les économistes préfèrent prendre tous les coûts et bénéfices en compte.

Une autre considération très importante lorsque les coûts et les bénéfices sont comparés à des périodes différentes est qu'il faut actualiser les coûts de façon à obtenir une évaluation correspondant à la réalité actuelle (Edejer *et al.*, 2003). L'Organisation mondiale de la Santé (OMS) recommande d'utiliser, dans l'analyse coût-efficacité, un taux d'actualisation standard de 3 % par an de la valeur actuelle pour des besoins de comparaison (voir l'étude de cas 2). Si on compare plusieurs pays, l'OMS recommande d'utiliser des dollars internationaux en se référant à des taux de change en parité de pouvoir d'achat (PPA) pour comparer les coûts (1 \$ permet d'acheter différentes quantités de produits de base dans des pays différents). Il est possible d'utiliser des outils spécialement conçus pour faciliter le calcul des coûts (voir, par exemple, le programme de CostIt).

L'évaluation des bénéfices d'une intervention dépend des informations sur l'efficacité, qui peuvent provenir des essais cliniques ou des données regroupées (par exemple, dans des revues systématiques). Il est nécessaire d'utiliser les modèles démographiques quand les interventions ont un effet quantifiable significatif sur l'espérance de vie. Ces modèles peuvent évaluer les effets sur la répartition de la population en fonction de l'âge et sur la fréquence spécifique des risques de maladie selon l'âge (Edejer *et al.*, 2003).

L'analyse de sensibilité est généralement utilisée pour mettre en évidence les incertitudes liées aux coûts ou aux bénéfices – en particulier lorsque la valeur des variables a dû être estimée. Dans l'analyse de sensibilité, un éventail de valeurs est introduit pour les paramètres incertains, afin d'évaluer la différence par rapport au ratio coût-efficacité final et aux résultats.

Les conclusions de l'analyse coût-efficacité doivent présenter de manière claire les informations suivantes de façon que les résultats puissent être interprétés et éventuellement appliqués dans d'autres contextes (Edejer *et al.*, 2003):

1. Les informations sur les coûts

- a. Comment les coûts unitaires et les quantités utilisées pour l'intervention ont-ils été calculés?
- b. Comment les coûts pour les patients ont-ils été évalués?
- c. Quelle est la source des données et de quelle manière les coûts ont-ils été répartis et actualisés?
- d. A-t-on utilisé les informations provenant des analyses de sensibilité ou d'autres analyses d'incertitude?

2. Les informations sur l'efficacité

- a. Est-ce qu'une revue systématique a été faite?
- b. Possède-t-on des informations détaillées sur les sources des données et les suppositions proposées? (Est-ce que l'efficacité a été modélisée et comment?)
- c. Connaît-on le nombre d'années de vie vécues par la population (les informations sur les scénarios utilisés dans l'étude)?
- d. Possède-t-on des informations sur les analyses de sensibilité ou d'autres analyses d'incertitude?

3. Les informations sur le ratio coût-efficacité

- a. A-t-on une représentation graphique des informations sur le ratio coût-efficacité en nombres ?
- b. A-t-on comparé le ratio coût-efficacité à l'« option nulle » pour toutes les interventions ?
- c. Les résultats des analyses d'incertitude sont-ils présentés sous forme de tableaux (facultatifs) ?

Quelques questions méthodologiques sur l'analyse coût-efficacité

Les bases et la méthodologie employées dans les évaluations économiques ont été abondamment examinées et longuement débattues (Weinstein et Manning, 1997; Koopmanschap *et al.*, 1995; Wagstaff, 1991; Gold *et al.*, 1996; Brouwer *et al.*, 2000). Une importante question d'ordre méthodologique concerne le choix de la mesure d'efficacité. Certains auteurs sont d'avis que l'analyse coût-efficacité devrait non seulement utiliser des effets de santé finaux, tels que les années de vie gagnées, comme une mesure d'efficacité, mais aussi considérer d'autres effets intermédiaires comme « le nombre de cas détectés ou le nombre de patients convenablement traités » (Drummond *et al.*, 1997, p. 98). Bien que l'utilisation des effets intermédiaires soit acceptée dans l'analyse coût-efficacité, il est important que celle-ci définisse clairement la relation causale et associe une valeur se situant entre l'effet intermédiaire et l'effet final. La mesure standard pour l'analyse coût-efficacité reste cependant l'effet final.

L'utilisation des données des essais cliniques dans l'analyse coût-efficacité peut aussi poser des problèmes méthodologiques, notamment lorsqu'on essaie de déterminer la comparabilité du contexte et des types de cas traités. En principe, des hypothèses peuvent également être formulées en l'absence de preuves cliniques fiables, mais à condition qu'une analyse de sensibilité soit effectuée par la suite.

D'autres débats ont porté sur l'utilité de considérer les coûts futurs liés aux soins de santé de même que ceux qui ne leur sont pas liés, l'inclusion des coûts de productivité, la prise en compte du principe d'équité, l'actualisation des coûts et des effets des interventions. D'autres auteurs se sont demandé si l'analyse coûts-efficacité devait être basée sur une approche visant le bien-être ou sur les besoins des décideurs (Brouwer et Koopmanschap, 2000). Selon la théorie du bien-être, les mesures comme

les années de vie ajustées par la qualité doivent être vues comme les mesures d'utilité basées sur le bien-être de l'individu, alors que le but de l'évaluation économique est de déterminer si un investissement fait dans une intervention sera bénéfique sur le plan social.

Les résultats de l'analyse coûts-efficacité peuvent être pondérés de façon à refléter les valeurs sociales (par exemple, donner plus de poids aux bénéfices qui touchent un groupe défavorisé). Cependant, cela devrait être fait d'une manière transparente pour que les résultats puissent être interprétés avec et sans pondération. Les évaluations des changements survenus dans le domaine financier ou dans les systèmes de santé sont particulièrement complexes (voir l'étude de cas 3).

Conclusion

L'évaluation économique est un outil utile qui permet d'informer les décideurs politiques sur les coûts et les bénéfices probables de différentes politiques ou interventions. Cependant, comme ce chapitre l'a montré, il est assez difficile de dégager ou d'interpréter les résultats des évaluations économiques. Il y a un certain nombre de débats méthodologiques en cours. De plus, les résultats d'une étude menée dans un contexte donné peuvent difficilement être appliqués dans un autre contexte, car les structures de coût varient et l'efficacité d'un programme dépend du contexte. Les mesures des effets ont aussi tendance à simplifier la réalité, la plupart des interventions ont un éventail large d'effets et ceux-ci ne sont pas tous faciles à évaluer quantitativement ou à inclure dans une mesure d'effet simple et comparable. Il serait peut-être judicieux de considérer l'évaluation économique non pas comme un outil permettant de répondre à des questions concernant la manière de prioriser et de répartir les ressources, mais plutôt comme un outil qui permet de mieux préciser ces questions.

ÉTUDE DE CAS 1

Une étude sur le rapport coût-efficacité des césariennes pratiquées par des agents intermédiaires, des médecins généralistes et des obstétriciens au Burkina Faso (Hounton *et al.*, 2009)

Cette étude avait pour but d'évaluer le coût-efficacité de l'utilisation des professionnels de santé alternative pour prendre en compte le problème de manque de professionnels qualifiés et améliorer l'accès aux soins obstétricaux d'urgence au Burkina Faso. Les données de 2006 ont été utilisées pour estimer les coûts de la formation et des ressources humaines.

Le ratio coût-efficacité incrémentiel a été calculé en divisant la différence entre les coûts moyens des équipes chirurgicales menées par les différents types de professionnels par la différence des taux de létalité des nouveau-nés associés à chaque option. Le coût moyen d'utilisation des services d'un obstétricien, d'un médecin généraliste ou d'une équipe chirurgicale dirigée par des agents intermédiaires a été divisé par le nombre moyen de césariennes réalisées. Les coûts variables ont été considérés comme étant les mêmes pour les équipes menées par les obstétriciens ou les autres équipes. D'autres coûts variables, tels que les coûts des kits qui sont surtout assumés par les patientes n'ont pas été inclus dans l'estimation étant donné que l'étude se plaçait au point de vue du prestataire des soins de santé.

Les résultats ont montré que les taux de létalité des nouveau-nés étaient plus élevés avec les agents intermédiaires qu'avec les obstétriciens et les médecins généralistes. Cependant, le coût par césarienne était en moyenne pour des obstétriciens de 92 858 CFA, pour les médecins généralistes de 37 531 CFA et pour les agents intermédiaires de 34 884 CFA. L'étude a permis de constater que si le nombre de césariennes réalisées devait augmenter, le coût par césarienne serait réduit significativement.

Malgré la difficulté de comparer les équipes de chirurgie dirigées par les obstétriciens, les médecins généralistes et les agents intermédiaires, en raison des différences de formation, de rémunération et des besoins d'assistance propres à chaque équipe, l'étude a montré qu'une formation alternative, en particulier la formation des médecins généralistes, conduit à une amélioration de l'accès aux soins obstétricaux d'urgence et est rentable. L'étude montre également qu'une meilleure planification des parcours professionnels et des primes pourrait améliorer la rentabilité de la formation des médecins généralistes. Les taux élevés de létalité avec les agents intermédiaires indiquent qu'une meilleure surveillance est nécessaire afin d'améliorer les soins administrés aux mères et de réduire le taux élevé de mortalité lorsque les césariennes sont pratiquées par ces équipes.

ÉTUDE DE CAS 2

Une analyse coût-efficacité de la distribution des moustiquaires imprégnées d'insecticide dans le cadre de la campagne intégrée de santé de l'enfant au Togo (Mueller *et al.*, 2008)

Le but de cette étude était d'évaluer la rentabilité de la fourniture de moustiquaires imprégnées d'insecticide à tous les enfants âgés entre neuf mois et cinq ans au Togo dans le cadre du programme national de vaccination contre la rougeole. Les coûts et les effets incrémentiels ont été calculés du point de vue du prestataire de soins. L'évaluation de l'efficacité de l'intervention reposait sur le nombre de cas de paludisme évités, sur le nombre de morts évitées et les années de vie d'incapacité évitées. Des évaluations régionales ont été utilisées pour prédire le nombre de cas de paludisme, et on a eu recours à des entrevues avec les informateurs clés et à des rapports financiers pour calculer le coût des traitements et des interventions.

L'approche incrémentale utilisée pour évaluer les coûts et les effets de l'introduction des moustiquaires imprégnées d'insecticide a été comparée à une « approche *statu quo* ». Les coûts de l'intervention comprenaient les coûts d'investissement, les coûts récurrents et les coûts d'opportunité. Comme l'intervention s'inscrivait dans le cadre du programme de vaccination contre la rougeole, certains coûts étaient partagés. Ils ont dû être répartis également.

Pour calculer les AVCI évitées, l'étude a utilisé les proportions évaluées par épisode des cas d'anémie, des séquelles neurologiques et le nombre de morts liées aux cas de paludisme. La durée de la maladie par épisode de paludisme a été évaluée comme étant de deux semaines, et les poids d'incapacité utilisés étaient ceux du projet de la charge mondiale de morbidité. Les AVCI ont été actualisées à 3 %, et une analyse de sensibilité a été effectuée pour les corrections sur les pondérations selon l'âge.

L'étude a permis de constater : 1) que le coût net d'un cas de paludisme prévenu était de 3,26 USD ; 2) que le coût net par mort évitée était de 635 USD et 3) que le coût net par AVCI était de 16,4 USD. Cette étude a considéré les coûts du point de vue des prestataires, donc les coûts et les bénéfices pour les bénéficiaires de l'intervention n'ont pas été pris en compte. Les économies de coûts pour les familles n'étaient pas incluses non plus dans cette étude. Ainsi, l'inclusion des coûts et des bénéfices pour les bénéficiaires augmenterait la rentabilité de tels programmes. L'étude met en évidence les bénéfices d'efficacité qui peuvent être obtenus en introduisant des campagnes de vaccination et des interventions de paludisme simultanément.

Ouvrages clés en évaluation économique

DRUMMOND, M., B. O'BRIEN, G. L. STODDART et G. W. TORRANCE (1997), *Méthodes d'évaluation économique des programmes de santé*, Paris, Economica.

Ce livre permet de se familiariser avec la théorie et la pratique de l'évaluation économique. Il décrit les méthodes de mesure des coûts, des résultats, de la collecte des données pendant les études cliniques et indique comment gérer l'incertitude dans l'évaluation économique.

EDEJER, T. T. et al. (2003), *Making Choices in Health: WHO Guide to Cost-Effectiveness Analysis*, Genève, World Health Organization.

Un guide qui présente les méthodes et les outils conçus par l'OMS.

WITTER, S., T. ENSOR, M. JOWETT et R. THOMPSON (2000), *Health Economics for Developing Countries: A Practical Guide*, Amsterdam, Royal Tropical Institute.

Si vous n'êtes pas économiste, voilà un guide qui vous expliquera les concepts de base d'une manière très simple, un guide centré sur les expériences des pays en voie de développement.

Références

AURAY, J. P., A. BERESNIAK, J. P. CLAVERANNE et G. DURU (1996), *Dictionnaire commenté d'économie de la santé*, Paris, Masson.

BROUWER, W. B. F. et M. A. KOOPSMAN (2000), « On the economic foundations of CEA. Ladies and gentlemen, take your positions! », *Journal of Health Economics*, 19, 439-459.

DELANDE, G. (1991), *Introduction à l'économie de la santé*, Montrouge, John Libbey Eurotext.

FARGEON, V. et P. SAVIGNAT (2009), *Introduction à l'économie de la santé*, Grenoble, Presses universitaires de Grenoble.

- FOX-RUSHBY, J. et J. CAIRNS (2005), *Economic Evaluation*, Angleterre, Open University Press.
- GOLD, M. R., J. E. SIEGEL, L. B. RUSSELL et M. C. WEINSTEIN (1996), *Cost-effectiveness in Health and Medicine*, Oxford, Oxford University Press.
- HOUNTON, S. H., D. NEWLANDS, N. MEDA et V. DE BROUWERE (2009), « A cost-effectiveness study of caesarean-section deliveries by clinical officers, general practitioners and obstetricians in Burkina Faso », *Human Resources for Health* 2009, (7)34, 1-12.
- HULL, R. D., J. HIRSH, D. L. SACKETT et G. L. STODDART (1981), « Cost-effectiveness of clinical diagnosis, venography and non-invasive testing in patients with symptomatic deep-vein thrombosis », *New England Journal*, 304, 1561-1567.
- KOOPMANSCHAP, M. A., F. F. H. RUTTEN, B. M. VAN INEVELD et L. VAN ROIJEN (1995), « The friction cost method for measuring indirect costs of disease », *Journal of Health Economics*, 14, 171-189.
- MAJNONI D'INTIGNANO, B. (2001), *Santé et économie en Europe*, Paris, Presses universitaires de France
- MORRIS, S., N. DEVLIN et D. PARKIN (2007), *Economic Analysis in Healthcare*, Chichester, UK, John Wiley and Sons.
- REYNOLDS, J. et K. C. GASPARI (1985), *Cost-Effectiveness Analysis*, Chevy Chase, Center for Human Services.
- ROBINSON, R. (1993), « Costs and cost-minimisation analysis », *British Medical Journal*, 307, 726-728.
- SCOTT, D. R. et D. S. SEAN (1999), « Weighing the economic evidence: Guidelines for critical assessment of cost-effectiveness analyses », *Journal of American Board Family Medicine*, 12, 6.
- SCULPHER, M. J. et M. J. BUXTON (1993), « The episode-free day as a composite measure of effectiveness », *Pharmacoeconomics*, (4)5, 345-352.
- SICURI, E. *et al.* (2010), « Cost-effectiveness of intermittent preventive treatment of malaria in pregnancy in Southern Mozambique », *PLoS ONE*, (5)10, 1-10.
- STIGLITZ, J. E., C. E. WALSH et J. D. LAFAY (2007), *Principes d'économie moderne*, Bruxelles, Éditions De Boeck Université.
- WAGSTAFF, A. (1991), « QALYs and the equity-efficiency trade-off », *Journal of Health Economics*, 10, 21-41.
- WEINSTEIN, M. C. et W. G. MANNING, Jr. (1997), « Theoretical issues in cost-effectiveness analysis », *Journal of Health Economics*, 16, 121-128.

12

Le renforcement des capacités en évaluation

Patricia Rogers et Marie Gervais

Ce chapitre traite de diverses questions et stratégies liées au renforcement des capacités en évaluation, c'est-à-dire la capacité des personnes, des organisations, des secteurs et des gouvernements à mener différents types d'activités de suivi et d'évaluation, puis à en utiliser les résultats. En premier lieu, nous présenterons le cadre conceptuel qui sous-tend notre réflexion sur le renforcement des capacités en évaluation, puis nous proposerons un éventail de moyens permettant de le stimuler, et ce, dans divers contextes de pratique.

La notion des capacités en évaluation

Une notion plus large que le développement d'habiletés

Nous entendons par « capacité d'évaluation » la capacité à mener des activités d'évaluation – soit le suivi continu et l'autoévaluation –, à assurer la gestion des évaluations externes, puis à en utiliser les résultats. Cette définition élargit délibérément le spectre des personnes et des groupes concernés (qui pourrait comprendre la société civile), la gamme des activités d'évaluation (au-delà des projets formels d'évaluation) et leur portée (on ne se limite pas à la réalisation d'évaluations, mais on valorise aussi

l'utilisation de leurs résultats). On pousse ainsi la notion de capacité au-delà de la simple possession d'habiletés et de connaissances pour englober la disponibilité des ressources ainsi que la présence d'une culture et de processus organisationnels favorisant leur utilisation.

Cependant, bien des gens définissent plutôt la capacité d'évaluation comme l'aptitude à effectuer une évaluation ponctuelle, un exercice ayant un début, un milieu et une fin, faisant l'objet d'un mandat formel et générant un rapport final. Milstein et Cotton (2000) ont ainsi rapporté, à la suite de la plénière présidentielle de la conférence annuelle de l'Association américaine d'évaluation, dont l'un des thèmes était « bâtir la capacité en évaluation », une définition de cette capacité comme étant « l'aptitude à effectuer une évaluation efficace, c'est-à-dire qui satisfasse aux normes acceptées dans le domaine ».

Bien qu'il soit maintenant reconnu que les habiletés, les ressources et les processus sont des éléments nécessaires au développement des capacités, ils ne sont toutefois pas suffisants pour bâtir et maintenir cette capacité en évaluation. À cet effet, Williams signale que « toutes les habiletés, connaissances, expertise technique et expérience au monde ne seront d'aucun secours si [...] le programme, la communauté, l'organisme ou l'environnement [...] ne [peuvent] maintenir et nourrir ces habiletés et aptitudes » (2001, notre traduction).

De même, les efforts de renforcement des capacités en évaluation sont souvent justifiés par le proverbe suivant : *Donne un poisson à un homme, et tu le nourris pour un jour; apprends-lui à pêcher, et tu le nourris pour la vie* (Fetterman, 1996, p. 11, lorsqu'il introduit l'idée d'évaluation visant l'habilitation). Par contre, si l'on reste dans le domaine de la métaphore, il ne suffit pas d'avoir les habiletés nécessaires pour attraper des poissons; il faut également un équipement et un réseau de distribution efficaces, des consommateurs et un système global de pêche viable. Bref, qu'il s'agisse de renforcer la capacité de pêcher ou d'évaluer, l'important est de travailler avec l'ensemble d'une organisation plutôt que de viser uniquement à perfectionner des habiletés individuelles.

Une notion qui concerne différentes instances

Le renforcement des capacités en évaluation peut se faire à l'échelle individuelle, organisationnelle, sectorielle et gouvernementale. Ainsi, les

efforts de développement des capacités d'évaluation se concentrent au niveau des individus (accroissement des habiletés), des organismes (procédures et développement des ressources) ou des groupements d'organismes. Quelquefois, ces efforts pour bâtir la capacité d'évaluation couvrent plus largement un système complet du secteur public.

À cela s'ajoute la capacité en évaluation de la communauté, par l'entremise de citoyens ou d'intervenants impliqués dans des évaluations participatives ou de consultants locaux engagés dans des évaluations offertes en sous-traitance. Il y a également le besoin de bâtir la capacité d'évaluateurs externes, en particulier là où leur nombre est relativement restreint.

Dans *The Encyclopedia of Evaluation*, Baizerman, Compton et Stockdill (2004, p. 38) proposent une définition du renforcement des capacités en évaluation pertinente à l'échelon organisationnel : « Le renforcement de la capacité en évaluation est un travail intentionnel visant la création et le maintien continus de processus organisationnels globaux qui mènent à une évaluation de qualité et à l'utilisation courante de ses résultats » (notre traduction).

Différentes orientations et stratégies sont manifestement nécessaires pour renforcer les capacités des personnes chargées des évaluations, de celles qui en utilisent les résultats ou encore de celles qui agissent au nom de la société civile.

Une notion qui comporte différents objectifs

Lorsqu'on cherche à développer les capacités en évaluation, il importe de cibler à quelle fin cette dernière sera utilisée, car les compétences jugées nécessaires pour mener avec succès les activités d'évaluation peuvent varier selon les caractéristiques du mandat d'évaluation et du contexte de pratique. Il faut donc avoir en tête la diversité des champs de pratique en évaluation ainsi que les connaissances et compétences particulières que doit posséder tout évaluateur.

Concrètement, il importe pour l'évaluateur d'élargir son éventail traditionnel d'activités et de ne pas se contenter d'évaluations ponctuelles. Les activités continues d'évaluation et de suivi deviennent ainsi autant d'éléments utiles et complémentaires pour rendre des comptes et viser une amélioration.

Toutefois, une consultation rapide de la documentation sur l'évaluation permet de constater qu'elle se concentre principalement sur un premier type d'activités, soit les évaluations ponctuelles. Ces dernières font souvent l'objet, à la fin d'un cycle de mise en œuvre, d'un rapport formel qui servira de base aux actions subséquentes. Elles alimentent ainsi les décisions relatives à la poursuite, à la reproduction ou à la cessation d'une initiative, d'un programme ou d'une politique, à l'allocation de ressources, à la reddition de comptes et au développement de connaissances servant à guider l'élaboration de nouveaux programmes, initiatives ou politiques reposant sur des données scientifiques. Ces évaluations peuvent être conduites par un organisme externe, par une unité d'évaluation interne distincte de l'unité de mise en œuvre ou encore par l'unité responsable de la mise en œuvre.

On constate de plus que la majorité des efforts de développement touchant les capacités en évaluation se concentrent également sur ce type d'activités d'évaluation. Ainsi, nombreux sont les organismes qui cherchent à développer les capacités de leur personnel (gestionnaires et praticiens) à gérer, à planifier et à réaliser des évaluations à l'interne, et ce, souvent dans une optique de complémentarité avec les évaluations octroyées à l'externe. Par ailleurs, il est maintenant fréquent de voir les ressources expertes externes sollicitées pour travailler avec une organisation afin de la guider et de l'outiller dans la mise en œuvre d'évaluations à réaliser à l'interne.

Il est aussi important de considérer un deuxième type d'activités dans le contexte du développement des capacités en évaluation, soit celles qui sont liées à l'évaluation continue au sein d'une initiative, d'un programme ou d'une politique. Réalisées sous un mode moins formel, ces activités peuvent prendre diverses formes, dont celles d'examen périodiques, d'inventaires, de rencontres ou de réflexions. Les activités d'évaluation continue peuvent aisément être utiles aux groupes locaux en leur permettant d'explorer leurs préoccupations, d'examiner des occasions de changements et de documenter les effets de ces changements. Habituellement, une organisation conserve des traces écrites de ses activités d'évaluation continue, ce qui facilite leur divulgation et favorise les discussions ouvertes sur les problèmes et les progrès à réaliser pour les résoudre. Les avantages considérables que peut retirer une organisation après la réalisation d'activités d'évaluation continue peuvent l'inciter à trouver des

modalités visant à soutenir concrètement ce type d'activités et à favoriser le développement de compétences particulières.

Les activités de suivi (ou monitoring) constituent le troisième type d'activités à considérer dans un contexte de développement des capacités en évaluation. Il faut viser le plus possible la formalisation de ces activités et encourager le développement de systèmes de suivi appropriés qui permettent de suivre dans le temps l'état ou l'évolution de certains objets d'évaluation. Par exemple, de tels systèmes permettent de détecter rapidement les écarts entre la performance affichée d'un indicateur et sa performance attendue, facilitant ainsi l'établissement de correctifs à mettre en place selon la situation en cause. Utilisés localement, ces types de systèmes peuvent également mener à des ajustements progressifs de la mise en œuvre d'une initiative, d'un programme ou d'une politique. En pratique toutefois, cette démarche demeure souvent assez difficile, spécialement en raison des délais de rétroaction. Il faut de plus porter une attention particulière à la qualité de tels systèmes. À titre indicatif, il importe d'éviter certains pièges comme utiliser des mesures inadéquates (pertinence non fondée des cibles de mesure, données de faible qualité et d'accès difficile), substituer de nouveaux objectifs à ceux qui ont été établis au départ, et cela en fonction des informations qu'on peut véritablement recueillir, insister sur des éléments anodins et accorder une importance exagérée aux moyennes.

Il est intéressant de constater que cette complémentarité entre les activités de suivi et les activités d'évaluation a été reconnue par de nombreuses associations et réseaux nationaux en plein essor, ceux-ci ayant choisi de centrer leurs activités sur le suivi et l'évaluation. La situation qui prévaut actuellement en Afrique et en Amérique latine en est un exemple manifeste.

Une notion qui s'actualise aux différentes étapes du processus évaluatif

Les capacités en évaluation doivent être présentes aux diverses étapes d'un processus évaluatif afin d'en assurer le succès (voir tableau 12.1). Il importe donc de veiller au développement de ces capacités dans chacune d'elles.

TABLEAU 12.1**Cibles de renforcement des capacités liées au processus évaluatif**

Étapes	Description
Le cadrage	Négocier et décider notamment quels seront les objectifs de l'évaluation, le public visé, les questions relatives à l'évaluation, les valeurs qui seront mises de l'avant, les ressources disponibles et les contraintes en présence (y compris les échéances).
La conception	Considérant le cadre de l'évaluation, décider comment répondre aux questions de l'évaluation et préciser la stratégie méthodologique (incluant l'échantillonnage, la collecte et l'analyse des données).
La conduite	Réaliser la mise en œuvre et appliquer la stratégie méthodologique retenue jusqu'à la production du rapport.
L'utilisation	Encourager l'utilisation de l'information à l'aide d'un ensemble de processus formels et informels, et dégager des enseignements du processus d'évaluation.
La gestion	Identifier et impliquer les principales parties prenantes et recourir aux processus courants de gestion de projets, incluant la gestion des étapes, la gestion financière, etc.

Il est primordial de reconnaître qu'une évaluation réalisée à l'externe ne peut porter que sur les rôles de conception et de conduite. L'organisation conserve ainsi la responsabilité des autres rôles, d'où la nécessité de posséder des connaissances et une expérience suffisantes en évaluation pour pouvoir cadrer, gérer et utiliser l'évaluation. Cela inclut par exemple le développement d'appels de propositions pour des services d'évaluation, l'étude, le choix et la gestion de ces propositions, et enfin l'examen et l'utilisation des rapports finaux.

Un cadre conceptuel pour le renforcement des capacités en évaluation

Quelques principes directeurs émanant de projets visant le renforcement des capacités des communautés

Le domaine communautaire a depuis quelques années déjà fait l'objet de projets intéressants visant le renforcement des capacités en évaluation. De ceux-ci, plusieurs enseignements ont pu être dégagés, lesquels peuvent maintenant servir à baliser de nouvelles expériences à réaliser dans d'autres milieux de pratique en évaluation. Ces projets reposent sur trois principes importants, principes qui viennent à leur tour s'appliquer au développement des capacités en évaluation :

- ils s'adressent à différents types de capitaux ;
- ils adoptent une approche s'appuyant sur les forces en identifiant et en consolidant les capitaux existants ;
- ils se concentrent sur le développement d'occasions d'utiliser des capacités qui ont été développées.

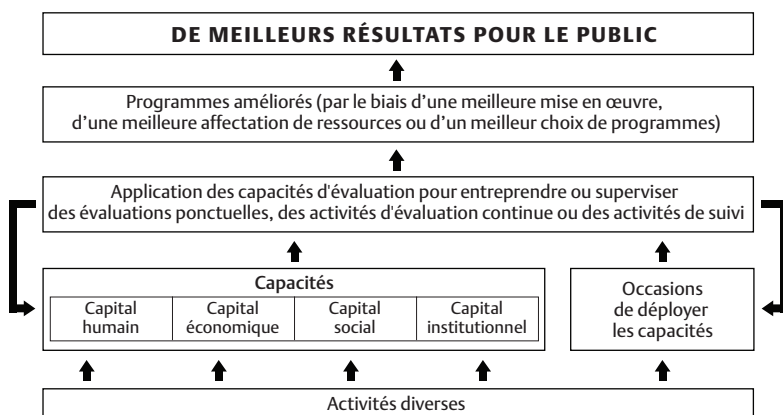
L'analyse de ces projets a permis de distinguer au moins quatre types de capitaux qui contribuent au développement des capacités des communautés (voir tableau 12.2).

TABEAU 12.2

Types de capitaux concernés par le renforcement des capacités en évaluation

Type de capital	Description
humain	Habilités en collecte et analyse de données, synthèse et élaboration de rapports, négociation, pensée évaluative.
institutionnel	Processus, structures et technologies qui soutiennent la production et l'utilisation d'informations pour guider les actions et décisions.
social	Normes de confiance et de réciprocité – confiance entre organisations et en leur sein pour permettre l'examen de questions difficiles, partager de l'information et explorer les doutes, incertitudes et échecs.
économique	Ressources adéquates sans réduction de services (au lieu de s'attendre à ce que les organisations entreprennent des activités d'évaluation significatives à même les budgets existants).

Ces principes directeurs conduisent naturellement au développement d'un cadre conceptuel permettant de baliser les composantes d'une démarche de renforcement des capacités en évaluation (voir figure 12.1). Ce cadre reconnaît quatre types de capitaux, alliant le développement de nouveaux capitaux au renforcement des capitaux existants ainsi qu'à la création par itération de possibilités d'utilisation des capacités développées et à leur mise à profit.

FIGURE 12.1**Cadre conceptuel sur le renforcement des capacités****Le capital humain**

Il y a actuellement un débat important sur les habiletés, connaissances et compétences essentielles requises pour mener des activités d'évaluation, et sur la nécessité de certifier ou d'accréditer les évaluateurs ou les cours de formation dans ce domaine. Malgré une tendance à une formation éclectique et à une évaluation utilisant une combinaison de méthodes, des différends demeurent au sujet des méthodes d'évaluation légitimes et essentielles qui, à ce titre, devraient faire partie des qualifications requises des personnes chargées d'une évaluation. Parmi elles, notons entre autres la collecte de données quantitatives ou qualitatives, les modèles expérimentaux et quasi expérimentaux ou naturalistes, les méthodes de collecte de données et d'analyse issues des sciences sociales ou encore les méthodes économiques. Il existe de plus des préoccupations quant au potentiel de généralisation de certaines compétences à différents contextes organisationnels et nationaux.

Dans ce contexte, les listes des compétences jugées essentielles en évaluation peuvent être très utiles, car elles constituent des références pour le perfectionnement continu de l'évaluateur et pour la création d'équipes d'évaluation disposant de connaissances et d'habiletés complémentaires. Il est cependant important de souligner que de telles listes portent sur les habiletés requises pour mener des évaluations à bien, et non sur l'utilisation de leurs résultats.

Une documentation significative a été produite au cours des dernières années au sujet de ces compétences essentielles. Le Canada et les États-Unis sont les chefs de file à ce chapitre, suivis de quelques pays européens. Examinons à titre indicatif la liste produite par Stevahn *et al.* (2005).

Les compétences essentielles d'après Stevahn

Pratique professionnelle 1. Application des normes et standards de pratique en évaluation 2. Respect du code d'éthique et recherche d'intégrité et d'honnêteté dans ses évaluations 3. Illustration des approches et habiletés aux clients potentiels 4. Respect des clients, répondants, participants aux programmes et autres parties concernées 5. Considération du bien général et public dans ses évaluations 6. Contribution au corpus de connaissances en évaluation Enquête systématique 1. Compréhension de la base de connaissances en évaluation (termes, concepts, théories et hypothèses) 2. Connaissance des méthodes quantitatives 3. Connaissance des méthodes qualitatives 4. Connaissance de méthodes combinant de multiples approches 5. Examen de la documentation 6. Précision de la théorie du programme 7. Formulation de questions d'évaluation 8. Élaboration d'un modèle d'évaluation 9. Précision des sources de données 10. Collecte des données 11. Examen de la validité des données 12. Examen de la fiabilité des données 13. Analyse des données 14. Interprétation des données 15. Établissement de jugements 16. Formulation de recommandations 17. Précision du raisonnement sous-jacent aux décisions prises lors de l'évaluation 18. Production de rapports touchant la démarche d'évaluation et les résultats 19. Identification des forces et des limites de l'évaluation 20. Tenue de méta-évaluations Analyse situationnelle 1. Description du programme 2. Détermination du caractère évaluable du programme 3. Précision des intérêts des autres parties concernées	4. Satisfaction des besoins des utilisateurs en matière d'information 5. Traitement des conflits 6. Examen du contexte organisationnel de l'évaluation 7. Analyse des considérations politiques pertinentes en rapport avec l'évaluation 8. Considération des questions liées à l'utilisation de l'évaluation 9. Considération des questions liées au changement organisationnel 10. Respect de l'individualité du site de l'évaluation et du client 11. Ouverture à la rétroaction 12. Modification de l'étude au besoin Gestion de projet 1. Réponse aux demandes de propositions 2. Négociation avec le client avant le début de l'évaluation 3. Rédaction d'ententes officielles 4. Communication avec le client pendant tout le processus d'évaluation 5. Budgétisation de l'évaluation 6. Justification des coûts relatifs aux besoins d'information 7. Identification des ressources nécessaires à l'évaluation, par exemple : informations, expertise, personnel et outils 8. Utilisation de la technologie appropriée 9. Supervision des autres personnes participant à l'évaluation 10. Formation des autres personnes jouant un rôle dans l'évaluation 11. Direction harmonieuse du déroulement de l'évaluation 12. Remise des travaux à temps Pratique réflexive 1. Conscience de soi en tant qu'évaluateur (connaissances, habiletés et dispositions) 2. Capacité réflexive sur sa propre pratique (compétences et aspects à améliorer) 3. Perfectionnement continu dans le domaine de l'évaluation

Source : Stevahn, L. *et al.* (2005), « Establishing essential competencies for program evaluators », *American Journal of Evaluation*, vol. 26, n° 1, p. 43-59. Notre traduction.

Le capital institutionnel

Le capital institutionnel comprend les processus, structures et infrastructures disponibles à des fins d'évaluation. La culture organisationnelle fait également partie du capital institutionnel. Davidson (2001) signale en effet que les politiques, les systèmes et les pratiques formels ne forment que « la partie visible de l'iceberg » de la culture d'apprentissage qui soutient ou inhibe l'évaluation ; les pratiques informelles, les actions symboliques, les croyances, les valeurs et les attitudes face à l'évaluation étant d'une égale importance.

Ainsi, parmi les actions symboliques importantes qui influencent la volonté des gens à participer à une évaluation, regardons par exemple ce qui arrive aux personnes qui prennent des risques, qui tâtent le terrain ou qui critiquent sévèrement. Sont-elles récompensées, encouragées ou pénalisées ? Certaines personnes sont-elles récompensées pour avoir atteint un objectif facile ou pour avoir pratiquement atteint un objectif difficile (Davidson, 2001) ? Quels sont à cet effet les systèmes de motivation, les éléments culturels et les actions symboliques qui viennent influencer une évaluation ? Cela oblige ceux d'entre nous qui œuvrent à l'amélioration des capacités en évaluation à travailler à la fois de haut en bas et de bas en haut au sein d'une organisation.

Le directeur d'un projet mis en œuvre au ministère des Industries primaires de Victoria, en Australie, et visant le renforcement des capacités en évaluation a pu observer ces bénéfices importants : « Cela a amené un changement de culture. C'est seulement maintenant que je prends conscience du fait que je fonctionne différemment. Nous pouvons maintenant tout naturellement remettre en cause ce que nous faisons. Il y a une atmosphère d'interrogation. Nous cherchons activement comment mieux faire. Nous n'avions pas cette attitude auparavant » (McDonald *et al.*, 2003, p. 21 ; notre traduction).

Le capital social

Le capital social a trait aux rapports entre les gens, plus particulièrement aux rapports de confiance et de réciprocité. Les personnes qui œuvrent au renforcement des capacités des communautés font la distinction entre le capital social affectif, qui lie des personnes se trouvant dans des situa-

tions semblables ou au sein d'une même communauté; le capital social relationnel, qui lie des personnes se trouvant dans des situations ou des communautés différentes; et le capital social instrumental, qui lie des personnes et des détenteurs de pouvoir, des ressources ou des sources d'information.

À titre d'exemple, le programme de perfectionnement International Program for Development Evaluation Training (IPDET), détaillé plus loin, vise clairement à développer non seulement le capital humain, sur le plan des connaissances et des habiletés, mais également le capital social, et cela sous la forme d'un réseau de soutien des évaluateurs.

Le capital économique

La présence d'un capital économique est un autre élément nécessaire au renforcement des capacités en évaluation. Ce capital vise à payer le personnel pour le temps consacré par le à l'apprentissage du processus d'évaluation, à la conduite d'évaluations et à l'utilisation des résultats qui en découlent. Il est également nécessaire pour obtenir de l'aide extérieure sous diverses formes (par exemple, formation, examen des évaluations internes et des systèmes d'évaluation, tenue d'évaluations externes), ou encore pour financer la formation, l'achat de livres, de revues, de logiciels ainsi que l'accès aux banques de données. Certains bailleurs de fonds inscrivent d'ailleurs l'évaluation comme condition à l'octroi de financement pour la réalisation de projets. Ainsi, Santé Canada exige de divers organismes qu'ils consacrent 10 % du budget de fonctionnement de leurs projets à la conduite d'une phase d'évaluation intégrée aux activités prévues.

Quelques moyens utiles pour le renforcement des capacités en évaluation

La formation

Différentes stratégies de développement professionnel peuvent être envisagées par l'évaluateur soucieux d'élaborer une pratique autonome et de qualité. Diversifiée et abondante, cette offre d'enrichissement des capacités en évaluation touche les différents types de capitaux énoncés précédemment.

Les programmes universitaires

Bien qu'elle soit encore limitée, la formation en évaluation progresse actuellement à un rythme notable à l'échelle internationale. Dans certaines universités, l'acquisition de connaissances et de compétences spécifiques à l'évaluation est formellement reconnue par un diplôme.

Cousins (2006), dans son examen pour le compte du Centre d'excellence en évaluation du Canada, a relevé 13 programmes de ce genre offerts au Canada, dont 3 en français. Il est question des programmes gradués de l'École nationale d'administration publique, de l'Université du Québec à Montréal et de l'Université Laval. Cette dernière offre d'ailleurs depuis septembre 2007 une spécialisation en évaluation au sein du programme de maîtrise en santé communautaire.

Cousins dresse également une liste de 15 formations diplômantes offertes à l'extérieur du Canada, soit 13 aux États-Unis, 1 en Australie et 1 autre au Royaume-Uni. Cette dernière est une initiative conjointe de l'Institute of Education de l'Université de Londres et du gouvernement. La formation proposée en Australie par le Centre for Program Evaluation de l'Université de Melbourne consiste en une formation accessible à distance aux étudiants de l'étranger et menant à une maîtrise en évaluation. L'Afrique anglophone compte également un certain nombre de programmes universitaires offrant une formation en évaluation appliquée à différents contextes de pratique.

Les formations abrégées

Un moyen dynamique pour renforcer ses capacités est de participer à des activités de perfectionnement professionnel. Il est question de conférences organisées annuellement par les réseaux professionnels en évaluation, d'ateliers de formation donnés tant à l'échelle nationale que régionale, et de cours abrégés s'échelonnant sur quelques jours ou quelques semaines.

Ainsi, l'IPDET, par l'entremise du Groupe d'évaluation de la Banque mondiale et de la Faculté des affaires publiques de l'Université Carleton, offre annuellement au Canada un programme de formation d'une durée d'un mois. Des établissements américains, dont les Centers for Disease Control et The Evaluators' Institute, proposent aussi des formations abrégées. L'Africa Sourcing, ainsi que le Centre africain des études supérieures en gestion de Dakar, au Sénégal, sont d'autres exemples d'organismes

offrant une formation en français. Ajoutons également la formation de quatre jours sur les compétences essentielles liées à l'évaluation offerte par la Société canadienne d'évaluation (SCE) et la Société québécoise d'évaluation de programmes (SQEP), en collaboration avec le Centre de liaison sur l'intervention et la prévention psychosociale (CLIPP).

Les programmes personnalisés avec pratique supervisée

Parmi les expériences probantes à ce sujet, mentionnons l'expérience réalisée par l'American Cancer Society (ACS), qui a soutenu le développement d'un programme de formation adapté aux besoins en évaluation du milieu médical. Cette société a notamment coordonné les partenariats nécessaires à l'échelle locale entre les agences, les universités, les collèges, les conseillers et les experts nationaux en évaluation, dont Michael Quinn Patton et Hallie Preskill, ce qui a permis aux étudiants d'entreprendre des évaluations portant sur les projets de l'ACS (Compton *et al.*, 2001).

L'apprentissage en ligne adapté au rythme de l'étudiant

La Banque interaméricaine de développement (IABD) utilise les technologies d'éducation à distance pour offrir de la formation sur le cycle d'un projet (de l'établissement et de la préparation d'un projet à sa gestion et à sa mise en œuvre, puis à son évaluation). Afin de former une masse critique de fonctionnaires en Amérique latine et dans les Caraïbes, celle-ci a conçu divers cours électroniques accessibles gratuitement sur son site Web.

Les documents de référence

À l'échelle internationale, on a recensé de très nombreuses initiatives visant l'élaboration de documents de référence destinés à outiller les évaluateurs œuvrant dans divers milieux de pratique. Ainsi, de nombreux partenaires pour le développement présentent sur leurs sites Web respectifs différentes informations utiles aux personnes qui désirent se perfectionner en évaluation. En plus de liens Internet pertinents et de rapports d'évaluation, on y trouve une abondance de documents qui abordent à la fois la théorie et la pratique en évaluation (lignes directrices, approches,

méthodes, outils) et qui traitent des bonnes pratiques, des grandes tendances et des enjeux en cours dans la profession. Le Fonds des Nations Unies pour l'Enfance (UNICEF), le Centre de recherches pour le développement international du Canada (CDRI), le Programme des Nations Unies pour le développement (PNUD) et la Fondation Kellogg en sont quelques exemples.

Il importe de mentionner que seule une faible proportion de cette documentation est rédigée en français, ce qui limite l'accessibilité à ce moyen de formation continue dans le monde francophone. De plus, sur le nombre toujours croissant de revues spécialisées et de bulletins publiés en évaluation à l'échelle internationale, seuls deux périodiques majeurs présentent des articles rédigés en français, soit la *Revue canadienne d'évaluation de programmes* et *Mesure et évaluation en éducation*.

Les normes et standards de pratique

Sur la scène internationale, le mouvement s'accroît concernant le développement ou la mise à jour des normes et standards de pratique en évaluation, que ce soit en Amérique du Nord, en Europe, en Océanie et, plus récemment, en Afrique. Leur prise en compte lors d'une démarche évaluative est donc fortement recommandée. On estime que ces normes et standards favorisent l'encadrement et le rehaussement de la qualité de la pratique évaluative (qualité technique et scientifique, conduite éthique et professionnelle), et qu'ils servent de support à la formation et à l'enrichissement professionnel. On les considère aussi comme des guides pour la méta-évaluation des expériences évaluatives réalisées et comme des éléments qui contribuent à la crédibilité et à la professionnalisation de ce champ de pratique. Ces normes et standards se trouvent sur les sites Web des principaux réseaux et associations professionnels de cette discipline.

Le soutien par les pairs et la pratique supervisée

Recourir à des pairs ou encore à des experts reconnus pour discuter des différentes facettes de son travail est un autre moyen de renforcer sa capacité en évaluation. Cela peut par exemple se concrétiser à l'intérieur d'un réseau plus ou moins formel de contacts, par des occasions d'échanges organisées par les réseaux professionnels en évaluation (par exemple, les

déjeuners thématiques de la Société wallonne d'évaluation), ou encore par la participation aux nombreux groupes de discussion en ligne destinés à la communauté des évaluateurs.

Dans une perspective de perfectionnement continu, il peut également être à propos de recourir à un superviseur ou à un mentor pour acquérir certaines expertises particulières en évaluation, que cela soit intégré ou non à des cours formels.

Les structures et processus

Le renforcement des capacités en évaluation s'effectue plus facilement lorsque cette fonction est soutenue par des structures, des processus et des politiques clairement définis. Cela s'avère un élément important dans le déploiement d'une culture d'évaluation au sein d'une organisation, ce qui constitue le plus souvent un passage obligé dans le processus de reddition de comptes imposé par une instance gouvernementale centrale. Cela renvoie toutefois à une des préoccupations actuelles dans ce domaine, soit la nécessité ou non de rendre l'évaluation de programmes obligatoire. En effet, il existe une crainte légitime voulant que cela puisse mener les évaluateurs à effectuer des évaluations qui ne seraient conformes qu'en apparence. Davies (1999, p. 157), par exemple, déclare catégoriquement que: « La leçon la plus importante est que la gestion du rendement (et l'exercice d'évaluation en général) ne peut pas être imposée aux gens. Tenter de l'imposer mènerait vraisemblablement à une substitution d'objectifs, à une information douteuse et à une augmentation du risque que la pertinence du programme en soit diminuée et non augmentée. »

Les réseaux professionnels de l'évaluation

L'évaluation est une discipline en plein essor dans le monde entier. Le nombre sans cesse croissant de nouveaux réseaux nationaux en évaluation (de 15 en 2000 à 61 en 2006) et la mise en place d'organisations-cadres comme l'International Organization for Cooperation in Evaluation (IOCE) et l'International Development Evaluation Association (IDEAS) sont devenus autant d'éléments qui contribuent à valoriser la discipline et à renforcer les capacités sur les plans local, national et international. L'IOCE a d'ailleurs publié en 2006 un livre sur les enseignements dégagés

des efforts de création de tels réseaux en Afrique, dans les Amériques, en Asie, en Australasie et en Europe.

En consultant régulièrement les sites Web de ces réseaux, on peut notamment être informé des occasions de perfectionnement (formations qualifiante et diplômante) et des événements professionnels à venir, ainsi qu'avoir accès à des banques de ressources fort utiles pour le renforcement des capacités (bibliothèques virtuelles, périodiques, standards de pratiques, rapports, livres, littérature grise, outils, groupes de discussions, etc.). Plus spécifiquement, en consultant le Portail de l'évaluation dans l'espace francophone, on peut repérer rapidement les réseaux francophones qui œuvrent aux quatre coins du globe et avoir accès à une banque importante d'informations et de ressources disponibles en français.

Mentionnons finalement les initiatives intéressantes conduites par la Société canadienne d'évaluation qui, chaque année, appuie le développement des capacités en évaluation chez la relève étudiante (bourses d'études universitaires, bourses de voyage permettant d'assister au congrès annuel, concours de simulation de cas, concours de meilleure dissertation).

Les prix de reconnaissance en évaluation

Plusieurs réseaux professionnels en évaluation ont choisi de décerner annuellement des prix d'excellence. Il est intéressant de noter qu'en plus d'être une manifestation tangible de reconnaissance par les pairs, ces prix ont démontré par le passé qu'ils ont encouragé le développement d'expertises et contribué au succès professionnel de leurs récipiendaires et à leur progression dans la carrière universitaire. À titre d'exemple, mentionnons l'Australasian Evaluation Society, qui a décerné en 2007 sept prix de reconnaissance, deux pour les contributions individuelles (nouveau talent et contribution remarquable) et cinq reconnaissant l'excellence d'organisations, d'équipes, de projets ou de programmes.

Le leadership assumé par une instance gouvernementale

Le gouvernement canadien s'est résolument tourné, depuis maintenant plus de 30 ans, vers le développement de la fonction d'évaluation comme moyen d'assurer une meilleure gouvernance. En 2001, la mise sur pied du Centre d'expertise en évaluation s'est avérée un levier important pour

appuyer le renforcement des capacités en évaluation, l'amélioration de la pratique et le renforcement de la communauté d'évaluateurs, et cela tant au sein du gouvernement fédéral que dans d'autres milieux de pratique au Canada. Le leadership assumé par cette instance s'est notamment exprimé par la multiplication d'initiatives visant à promouvoir l'évaluation comme fonction essentielle au processus de gestion, à soutenir les évaluateurs quant aux défis liés à leur pratique et à produire une importante documentation ayant trait à cette pratique.

Conclusion

Compte tenu des multiples objectifs de l'évaluation et de la diversité des connaissances et des compétences requises pour la prestation d'activités de qualité, et considérant que l'approfondissement des capacités en évaluation nécessite différentes ressources et occasions de perfectionnement, il est reconnu qu'un investissement dans ce domaine constitue une démarche complexe qui doit être soutenue au fil du temps. Ainsi, peu importe le niveau où existe une volonté de renforcer ces capacités (individuel, organisationnel, systémique), il est nécessaire d'adopter une démarche proactive, dynamique et étapiste. Quelques facteurs clés ont été dégagés à cette fin de certaines expériences de renforcement des capacités.

Tout d'abord, on recommande d'avoir une planification claire des objectifs visés en termes d'accroissement des capacités et d'avoir les moyens qui correspondent à ses ambitions. De plus, quand on pense par exemple aux changements qui peuvent survenir dans l'environnement sociopolitique d'une évaluation et aux changements de pratiques qu'ils entraînent nécessairement, ou encore au roulement de personnel qui se produit souvent dans ce champ de pratique, il devient primordial d'avoir une vision de renforcement des capacités qui déborde d'un horizon de court terme.

Nous suggérons également d'adopter une politique de « petits pas » et de réaliser des expériences pilotes avant de se lancer dans de grands chantiers. Cela permet de construire à la fois sur des expériences conduites avec succès et sur des acquis manifestes. Ces facteurs viennent graduellement contribuer à la croissance d'une culture d'évaluation et au lien de confiance qu'il est important d'établir face au processus, et ce, surtout lorsqu'il y a des craintes, souvent bien légitimes, concernant la mise en relief éventuelle de problèmes ou de failles de rendement.

À cela s'ajoute la nécessité d'investir non seulement dans la formation et l'acquisition de connaissances, mais aussi dans la supervision ou encore l'accompagnement des nouveaux évaluateurs dans la réalisation même d'activités d'évaluation. Il s'agit ici de favoriser le transfert de la théorie vers la pratique. Dans le même ordre d'idées, il importe également de soutenir l'accès à de la formation de plus en plus experte au fil du temps, et d'appuyer concrètement les changements de pratiques générés par l'acquisition de ces nouvelles expertises.

Un autre élément à considérer dans l'accroissement des capacités en évaluation est l'obtention, le plus tôt possible dans le processus de renforcement, de résultats concrets et pouvant être objectivés. Et de là, il importe que l'évaluation soit perçue comme répondant notamment à des besoins à court terme, ce qui contribue à son utilisation immédiate (McDonald *et al.*, 2003). Allant dans le même sens, et s'appuyant sur les travaux de Patton, Ridde et Sahibullah (2006) ajoutent qu'une approche de l'évaluation axée sur l'utilisation a été un élément clé ayant favorisé la participation des gens aux activités de renforcement des capacités en évaluation en Afghanistan. Poursuivant leurs réflexions, ces auteurs ajoutent qu'il leur a été utile d'apprendre à saisir les différentes occasions qui se sont présentées à tout moment du processus évaluatif pour travailler au renforcement des capacités en évaluation.

Finalement, compte tenu de l'essor dynamique de cette discipline et du fait que les gestionnaires y ont recours de manière nettement plus systématique pour alimenter la prise de décision et éclairer l'action, il est devenu impératif d'investir de diverses manières dans la création de ressources expertes afin de répondre à cette demande sans cesse croissante.

De fait, l'accélération des efforts de renforcement des capacités en évaluation constatée tant à l'échelle locale qu'à l'échelle internationale, la multiplicité des organismes et des personnes mobilisées dans un tel processus ainsi que l'abondance de la documentation de qualité publiée pour l'appuyer sont autant d'indications notables que le renforcement des capacités en évaluation est maintenant perçu comme étant l'affaire de tous. En ce sens, cela interpelle à la fois la responsabilité individuelle de chacun et la responsabilité collective de la communauté des professionnels.

Pour aller plus loin :

HORTON, Douglas *et al.* (2004), *L'évaluation au cœur du renforcement organisationnel. Expériences d'organisations de recherche et développement du monde entier*, ISNAR/CRDI/CTA.

À partir des résultats du Projet d'évaluation du renforcement des capacités lancé en 2000, au sein duquel six évaluations ont été réalisées sur divers continents sur une période de trois ans, cet ouvrage s'attarde à montrer en quoi les initiatives de renforcement des capacités locales et les évaluations périodiques permettent aux organisations de recherche et de développement de remplir leur mandat.

MCDONALD, Bron, Patricia J. ROGERS et Bruce KEFFORD (2003), « Teaching People to Fish? Building the Evaluation Capability of Public Sector Organizations », *Evaluation*, vol. 9, n° 1, p. 9-29.

À partir d'une étude de cas réalisée au sein d'une agence gouvernementale, cet article se penche sur les éléments qui ont contribué au renforcement des capacités en évaluation en analysant chacune des phases de son implantation réalisée sur une période de cinq ans. Il présente aussi plusieurs recommandations issues des discussions menées avec le personnel et les gestionnaires impliqués dans le projet.

PRESKILL, Hallie et Darlene RUSS-EFT (2005), *Building Evaluation Capacity: 72 Activities for Teaching and Training*, Thousand Oaks, Sage Publications.

Cet ouvrage de référence, principalement destiné à des activités d'enseignement et de formation, couvre l'ensemble des étapes nécessaires à la conduite d'une démarche d'évaluation, des fondements de cette discipline à la création d'un environnement favorable à son utilisation, en passant notamment par la collecte et l'analyse de données. Chacun des thèmes abordés est accompagné d'une série d'exercices permettant aux participants d'intégrer et de maîtriser leur contenu.

Internet

Centre de recherches pour le développement international (CRDI).

La section Évaluation du CRDI fournit une vaste documentation sous forme de liens menant vers des publications et des ressources spécialisées en évaluation, incluant notamment des lignes directrices, des livres, des rapports et d'autres formes d'outils; des évaluations stratégiques comprenant des ressources telles que le renforcement des capacités, les réseaux, l'influence de la recherche sur les politiques ainsi que le projet des subventions par concours; l'apprentissage organisationnel, qui présente des informations sur les rapports de fin de projet dynamique (RFPd), un outil permettant de tirer des enseignements des projets réalisés et de les diffuser, et sur le Forum annuel sur

l'apprentissage; des articles et des présentations sur l'évaluation et une section présentant les activités, les projets de même que l'équipe chargée de l'évaluation au CRDI. www.idrc.ca/fr/ev-26266-201-1-DO_TOPIC.html

Portail de l'évaluation dans l'espace francophone.

Ce site destiné aux professionnels propose en français des liens commentés vers des ressources en plusieurs langues sur l'évaluation, sur ses méthodes et sur des événements précis. Financé par l'Organisation internationale de la francophonie [OIF], il comprend environ 500 articles sur les thèmes suivants : réseaux professionnels et autres organisations contribuant au développement de l'évaluation, manifestations professionnelles à venir ou antérieures, formations longues ou courtes, croissance de l'activité professionnelle, méthodes et terminologie, autres ressources documentaires sur l'utilité et la déontologie de la discipline, et bases de données de rapports d'évaluation en ligne. Il est également possible d'accéder aux archives de ce site, de même qu'aux articles qui concernent un pays ou un continent en particulier, ou encore un secteur d'activité précis. www.evaluation.francophonie.org

Le site de la Société canadienne d'évaluation (SCÉ).

En plus de présenter la Société, son organisation, son congrès annuel et la *Revue canadienne d'évaluation de programmes* qu'elle publie, et de rendre accessibles des bulletins électroniques, ce site présente différents services offerts comme des concours étudiants, des prix et des offres d'emplois, de contrats et de collaboration. Il brosse également un tableau des diverses possibilités de perfectionnement professionnel offertes au Canada et ailleurs dans le monde, notamment sous la forme de congrès scientifiques, de formations spécialisées et de cours postsecondaires. Les ressources disponibles sont nombreuses et regroupées sous divers thèmes : les avantages, extraits et connaissances liés à l'évaluation, une bibliographie canadienne, la littérature grise, une liste de diffusion, des sections sur l'éthique et sur la coopération internationale, une banque de rapports d'évaluation, une liste de liens pertinents et de groupes de discussion, ainsi que des suggestions d'ouvrages de référence. Ce site permet également d'accéder directement aux 12 sections régionales de la SCÉ. www.evaluationcanada.ca

13

Conceptualiser et mesurer la participation à l'évaluation*

Pierre-Marc Daigneault et Steve Jacob

L'implication des parties prenantes (*stakeholders*) dans l'évaluation est un principe généralement admis dans le monde de l'évaluation depuis plusieurs années déjà (Whitmore, 1998). La participation est en effet devenue un concept central dans le domaine (Fleischer et Christie, 2009 ; Mark, 2001), voire un concept avec lequel « [...] chaque évaluateur et théoricien de l'évaluation est d'accord » (Mathison, 2005b, p. xxxiii : traduction). Le consensus concernant les approches et les techniques participatives est si fort que certains l'ont comparé à un « acte de foi » (Shea et Lewko, 1995, p. 159) ou l'ont qualifié d'« orthodoxie participative » (Biggs,

* Ce chapitre constitue une version revue, mise à jour et abrégée de l'article suivant : Daigneault, P.-M. et S. Jacob (2009), « Toward accurate measurement of participation: Rethinking the conceptualization and operationalization of participatory evaluation », *American Journal of Evaluation*, 30(3), 330-348. La section sur les principes et les outils utilisés pour analyser le concept de l'évaluation participative (ÉP) a été substantiellement réduite. Nous renvoyons le lecteur que le sujet intéresse à l'article en question. Des modifications à la conceptualisation et à son opérationnalisation ont également été apportées. Par souci de clarté, nous avons souligné ces modifications dans des notes de bas de page et en conclusion. Nous remercions le Conseil de recherches en sciences humaines du Canada (CRSH) et l'École de la fonction publique du Canada (EFPC) pour le soutien financier apporté à nos travaux sur l'évaluation participative.

1995, cité dans Gregory, 2000, p. 180). De là à soutenir que l'évaluation participative ne comporte que des avantages ou est appropriée en toutes circonstances, il n'y a qu'un pas que tous ne sont pas prêts à franchir (Gregory, 2000; Jacob et Daigneault, 2011; Jacob et Ouvrard, 2009). À l'inverse, quelques auteurs sont beaucoup plus critiques à l'égard des démarches participatives, ils dénoncent l'amateurisme et les menaces que font peser ces approches sur la validité, la crédibilité et l'éthique de l'évaluation (Scriven, 1997, 2005). Cela étant dit, même ceux qui ne considèrent pas la participation comme un idéal ne peuvent guère contester, dans le domaine de l'évaluation, la pertinence de thèmes comme la participation, l'inclusion et l'habilitation des parties prenantes.

Si la popularité de l'ÉP est une bonne nouvelle pour ses défenseurs, elle occasionne des problèmes sur le plan conceptuel. Comme l'affirme Huberman (1995) : « L'évaluation participative est un concept noble, mais insaisissable, qui semble refaire surface tous les 30 ans sous une autre forme rhétorique, mais qui pose toujours les mêmes problèmes conceptuels et pratiques épineux » (p. 104 : traduction). Plusieurs se sont d'ailleurs rangés à cet avis (Murray, 2002; Rebien, 1996; Ridde, 2006). Cette situation peu enviable freine le développement des connaissances théoriques et empiriques sur les approches participatives et leurs conséquences, et elle nuit en outre à la communication entre commanditaires de l'évaluation et évaluateurs. Cela est particulièrement préoccupant étant donné les appels répétés en faveur d'un accroissement des recherches sur l'évaluation en général et sur l'ÉP en particulier (Christie, 2003; Cousins, 2001; Cousins et Earl, 1999; Mark, 2001).

Ce chapitre tente de remédier aux problèmes précédents en proposant une analyse systématique et une opérationnalisation cohérente du concept d'ÉP. L'objectif poursuivi est double. Sur le plan conceptuel, il s'agit de circonscrire les attributs fondamentaux de l'ÉP afin d'en améliorer la clarté sémantique et la cohérence. Sur le plan empirique et pratique, il s'agit de trouver des indicateurs précis de l'ÉP afin de permettre au chercheur et au praticien de classer les évaluations selon leur niveau de participation et de distinguer efficacement les cas participatifs et non participatifs.

Après un bref examen des défauts conceptuels de l'ÉP, nous proposons une conceptualisation théoriquement fondée qui s'appuie sur le cadre d'analyse conçu par Cousins et Whitmore (1998). Nous déterminons ensuite des indicateurs précis afin de rendre la conceptualisation opéra-

tionnelle. Enfin, un cas d'application concret montre l'utilité du cadre conceptuel.

Un concept imparfait

Un concept est traditionnellement considéré comme une entité symbolique faisant le pont entre le monde réel et la pensée (Schedler, 2011). On représente généralement le concept sous la forme d'un triangle dont les angles sont le *terme* (le mot utilisé pour le désigner), le *sens* (sa définition) et le *réfèrent empirique* (la catégorie d'objets auxquels il renvoie) (Daigneault et Jacob, 2012; Sartori, [1984], 2009). La valeur d'un concept est fonction de la relation entre ces éléments et de critères tels que la clarté, la précision, le pouvoir de différenciation, la parcimonie et la cohérence interne (Gerring, 1999; Goertz, 2006).

Le concept d'ÉP présente plusieurs défauts. On note tout d'abord une ambiguïté dans la relation entre le terme et le sens du concept. Un concept clair (c'est-à-dire non ambigu) possède en effet un seul terme et un seul sens dans un contexte donné. La multiplication des termes employés pour désigner les différentes approches évaluatives impliquant les parties prenantes tels que « centrée sur l'utilisation », « pluraliste », « démocratique », « collaborative » et « de quatrième génération » (King, 1998), témoigne de ce problème. Par ailleurs, les définitions de l'ÉP abondent et elles sont souvent complexes et contradictoires. Burke (1998) cite par exemple jusqu'à sept principes de l'ÉP et autant d'éléments caractérisant son processus. Étant donné la multiplication des termes et des attributs, il est assez difficile de s'y retrouver. Ensuite, la plupart des conceptualisations ne sont pas opérationnalisées. En d'autres termes, il est difficile d'appliquer avec précision ces conceptualisations à des cas réels, par exemple pour différencier un cas d'ÉP d'un cas d'évaluation non participative ou encore pour classer les différents cas selon leur niveau de participation.

Refonder le concept d'évaluation participative sur des bases solides

Afin d'effacer toute ambiguïté et toute imprécision, nous proposons de refonder le concept d'ÉP sur des bases théoriques plus solides. Il faut, d'une part, utiliser les principes, les critères et les outils d'analyse conçus par Gerring (1999) et Goertz (2006) pour améliorer la structure de ce

concept, et, d'autre part, définir les attributs fondamentaux de l'ÉP en se référant aux écrits pertinents du champ de l'évaluation.

Un cadre d'analyse porteur

S'inspirant de travaux portant sur le sujet (Cousins, Donohue et Bloom, 1996; Cousins et Earl, 1992, 1995), Cousins et Whitmore (1998) ont ciblé ce qui nous apparaît être les attributs fondamentaux de l'ÉP. Leur cadre d'analyse établit une distinction entre deux « courants » d'ÉP, l'évaluation *pratique* et l'évaluation *transformative*, mais ce sont d'abord les trois dimensions proposées pour caractériser l'ÉP qui retiennent notre attention. La première dimension est le *contrôle du processus évaluatif*, qui renvoie au contrôle des décisions techniques liées à la conduite de l'évaluation (par opposition aux décisions concernant l'opportunité d'amorcer une évaluation). Elle couvre un spectre allant d'un contrôle total exercé par l'évaluateur à un contrôle total exercé par d'autres parties prenantes. La deuxième dimension, la *sélection des parties prenantes* qui participent à l'évaluation, a rapport aux types de parties prenantes impliquées dans l'évaluation. Elle s'étend de l'inclusion des principaux utilisateurs (*primary users*) à celle de tous les groupes légitimes. La troisième dimension est la *profondeur de la participation* (*depth of participation*). Elle s'étend de la consultation (sans pouvoir décisionnel) à l'implication dans tous les aspects de l'évaluation (conception, collecte et analyse des données, diffusion et utilisation des résultats) (Cousins et Whitmore, 1998).

Nous considérons que les trois dimensions précédentes sont des attributs fondamentaux, *nécessaires et suffisants*, de l'ÉP (voir Goertz, 2006). D'une part, chacune de ces trois dimensions doit être présente (relation de nécessité) pour qu'on ait le concept de participation. D'autre part, aucune autre dimension n'est requise pour qu'on puisse affirmer qu'on est en présence de la participation (relation de suffisance)¹. Ainsi unies par la logique des conditions nécessaires et suffisantes, les dimensions établies par Cousins et Whitmore (1998) permettent de distinguer efficacement

1. En 2004, Weaver et Cousins ont modifié le cadre original en ajoutant deux dimensions: facilité de la mise en œuvre de l'évaluation (de gérable à ingérable) et rapports de pouvoir entre les parties prenantes (de conflictuels à neutres). Ces changements ont une utilité sur le plan analytique, mais cependant ils ne peuvent être considérés comme des attributs fondamentaux et nécessaires du concept d'ÉP.

l'ÉP d'autres approches évaluatives (voir le chapitre 1). Nous croyons par ailleurs que ce cadre correspond plus ou moins explicitement aux intuitions qu'ont la plupart des évaluateurs, comme en témoigne la vaste utilisation du cadre à l'intérieur et à l'extérieur du domaine de l'évaluation (par exemple, Butterfoss, Francisco et Capwell, 2001; King, 2007; Themessl-Huber et Grutsch, 2003).

Plusieurs définitions de l'ÉP sont compatibles avec les dimensions proposées par Cousins et Whitmore (1998). Mathie et Greene (1997) ont par exemple donné la définition suivante: « Une caractéristique définissant l'évaluation participative est l'engagement actif de plusieurs parties prenantes » (p. 279: traduction). Rossi, Lipsey et Freeman (2004) ont proposé une définition analogue: « Les parties prenantes participantes interviennent directement dans la planification, la réalisation et l'analyse de l'évaluation en collaboration avec l'évaluateur, qui sera peut-être un chef d'équipe, un spécialiste ou une personne consultée au besoin seulement » (p. 51: traduction). De même, King (2005) a fait observer que l'ÉP « [...] est un terme générique qui désigne toute approche d'évaluation faisant intervenir activement le personnel chargé des programmes ou les participants dans le processus de décision et dans d'autres tâches de planification et de mise en œuvre des activités d'évaluation » (p. 291: traduction).

Ces définitions mettent toutes en évidence le fait que les parties prenantes non évaluatives à l'évaluation (c'est-à-dire les participants autres que l'évaluateur) *interviennent directement dans la réalisation de l'évaluation*. Par contre, dans l'évaluation non participative, les parties prenantes ne sont pas mises à contribution et, lorsqu'elles le sont, leur implication est indirecte et elle se limite souvent à la planification du processus évaluatif ou à la fourniture de données. La participation se définit au contraire par le fait que les parties prenantes sont réellement des *sujets*, et non des objets, de l'évaluation, qu'elles exercent un certain contrôle sur des aspects tels que la conception de l'évaluation (King, 2005). L'importance de la dimension du contrôle du processus est d'ailleurs reconnue dans les études consacrées à l'évaluation (Burke, 1998; Greene, 1987, 2005; King, 2005; Murray, 2002; Rebien, 1996; Weiss, 1986).

Renforcer la conceptualisation de l'évaluation participative

Le cadre conceptuel de Cousins et Whitmore (1998) représente un progrès marqué par rapport aux conceptualisations précédentes. Il présente cependant quelques lacunes auxquelles il faut remédier pour pouvoir l'utiliser comme instrument de mesure.

En premier lieu, le cadre ne permet pas le classement des cas d'évaluation sur la base de leur *niveau général* de participation. Si différents cas peuvent être comparés quant à la diversité des participants, la profondeur de la participation et au contrôle du processus évaluatif, le cadre n'offre aucune échelle de mesure permettant de déterminer si une évaluation est plus participative qu'une autre ou si elle l'est moins. De plus, le cadre n'indique pas quelle est la structure qui unit les dimensions au concept général de l'ÉP. Il importe donc de créer un indice qui permette aux chercheurs de mesurer avec précision la participation globale et d'étudier la relation de celle-ci avec d'autres variables comme l'utilisation de l'évaluation.

On constate en second lieu l'absence d'un « point zéro » explicite et cohérent pour le concept général et les dimensions de diversité et de profondeur. Le point zéro n'est rien d'autre que l'absence logique du concept ou de la dimension (voir Goertz, 2006, p. 32). Il est par ailleurs nécessaire d'établir une démarcation entre la présence (pôle positif) et l'absence (pôle négatif) du concept de base et de ses dimensions. En ce qui a trait au concept de base, l'ÉP est le pôle positif du continuum, tandis que l'évaluation non participative en est le pôle négatif.

Les pôles négatifs sont à revoir pour les dimensions *diversité* et *profondeur* (la dimension *contrôle* possède un point zéro explicite et cohérent). La dimension diversité va de l'inclusion des utilisateurs primaires dans le processus évaluatif à tous les groupes légitimes. Or, une évaluation non participative comprend au moins un commanditaire de l'évaluation et un évaluateur; c'est le pôle négatif de la dimension. Quant à la dimension profondeur, sa conceptualisation est défectueuse, car les indicateurs utilisés à chaque pôle réfèrent à des concepts différents. À son pôle positif, la dimension profondeur est clairement conceptualisée comme l'implication des participants dans toutes les tâches d'une évaluation, à savoir la conception, la collecte de données, l'analyse, la rédaction de rapports, la diffusion et l'utilisation des résultats. La *consultation* est vue comme le pôle négatif (voir cependant Cousins, Donohue et Bloom, 1996, p. 209-211).

Or, l'opposé de l'implication dans *toutes* les tâches d'évaluation est l'implication dans *aucune* tâche ou, autrement dit, l'absence d'implication dans ces tâches. On peut en effet imaginer une évaluation où les parties prenantes sont impliquées dans toutes les étapes du processus mais sont simplement consultées (elles ont alors une influence limitée). Puisque le terme consultation a rapport à l'influence des participants sur le processus plutôt qu'à leur implication dans celui-ci, le pôle négatif de cette dimension semble « contaminé » par le contrôle. Comme le contrôle possède déjà sa propre dimension, cela peut entraîner de la confusion ou un double comptage.

En dernier lieu, Cousins et Whitmore (1998) n'ont donné que peu d'indications concernant la manière d'appliquer concrètement et avec précision leur cadre. À partir de quel point sur une dimension peut-on considérer une évaluation comme participative ? Leur conceptualisation doit donc être d'abord opérationnalisée pour qu'elle puisse servir à mesurer l'ÉP. Cela implique qu'on convertisse les dimensions théoriques en indicateurs et en procédures concrètes permettant d'observer empiriquement le concept.

Les deux sections qui suivent s'attachent à corriger ces lacunes et proposent une version améliorée du cadre de Cousins et Whitmore (1998).

De la conceptualisation à la mesure

Étendue de l'implication

Nous avons remplacé l'expression *profondeur de la participation* par celle d'*étendue de l'implication* car elle donne une meilleure idée du rôle joué par les parties prenantes dans certaines tâches d'évaluation. Le terme *étendue* rend mieux compte du nombre de tâches accomplies par les parties prenantes que le terme *profondeur*, qui a plutôt rapport à la qualité et à l'intensité de l'implication. Comme l'évaluation participative concerne l'implication des parties prenantes dans l'élaboration de l'évaluation, nous limitons la mesure de cette caractéristique aux tâches *techniques* du processus évaluatif telles que la conception de l'évaluation et l'interprétation des résultats (voir Weaver et Cousins, 2004). Le terme « technique », montre bien que ces tâches sont liées étroitement à la production de l'évaluation, ce qui n'est pas le cas de la décision d'entreprendre une évaluation

ou de l'utilisation des résultats. Les tâches techniques sont habituellement considérées comme relevant de l'évaluateur². Toute évaluation comporte les étapes ou points de décision suivants :

1. La *conception de l'évaluation* est l'étape où l'on définit les questions et enjeux de l'évaluation et où on élabore le devis d'évaluation, où on choisit le cadre théorique, les méthodes, les techniques et les instruments.
2. La *collecte et l'analyse des données* est l'étape où l'on collecte, organise, code et analyse les données (données administratives, entrevues, etc.).
3. La *formulation des jugements et des recommandations* est l'étape où l'on porte un jugement sur la valeur d'un programme en s'appuyant sur les données et les résultats enregistrés à l'étape précédente, d'une part, et où l'on formule des pistes d'action plus ou moins formelles, de l'autre.
4. La *présentation et la diffusion des résultats de l'évaluation* correspondent à l'étape où l'on dégage et où l'on diffuse les résultats, verbalement ou par écrit.

Chaque tâche est considérée comme un indicateur dichotomique (oui/non) de l'*implication des parties prenantes non évaluatives*. Rappelons que les parties prenantes non évaluatives sont définies comme celles ayant un intérêt dans un programme ou son évaluation, sans être des évaluateurs. Pour qu'il y ait implication dans une tâche, les parties prenantes non évaluatives doivent jouer un rôle *significatif* dans la réalisation de celle-ci (c'est-à-dire être des acteurs et non pas seulement des fournisseurs de données)³. Aucun indicateur précis (p. ex., implication dans une tâche particulière comme la conception de l'évaluation) n'est nécessaire pour qu'on considère qu'il y ait implication. Autrement dit, les indicateurs peuvent se substituer les uns aux autres, ce qui signifie que toutes les tâches ont la même importance.

2. Contrairement à Cousins et Whitmore (1998), nous ne croyons pas que l'utilisation des résultats d'une évaluation soit une étape de l'évaluation à proprement parler, si souhaitable que cela puisse être.

3. L'opérationnalisation originale de cette dimension (voir Daigneault et Jacob, 2009) impliquait que la simple présence des parties prenantes à une étape de l'évaluation est suffisante pour qu'il y ait participation. Or, une étude empirique de validation de l'instrument de mesure (Daigneault, Jacob et Tremblay, en préparation) nous a permis de constater que cette règle pouvait sérieusement compliquer le codage et ainsi entraîner des erreurs.

Une ligne de démarcation est toutefois nécessaire pour pouvoir distinguer les cas où il y a implication de ceux où il n'y en a pas. Nous sommes d'avis que la participation à *une* tâche est le minimum requis pour qu'une évaluation soit considérée comme participative du point de vue de l'étendue de l'implication. Le tableau 13.1 contient les valeurs de codage représentant le niveau d'appartenance à cette dimension. Chaque indicateur a une pondération de 0,25. Cette formule simple facilite l'interprétation de l'étendue de l'implication ainsi que la combinaison avec les autres dimensions. L'idée qui sous-tend le codage de cette dimension est que plus le nombre de tâches dans lesquelles interviennent les parties prenantes non évaluatives est élevé, plus l'évaluation est participative, toutes choses étant égales par ailleurs.

Il faut préciser que, passé un certain niveau, le nombre de parties prenantes et leur niveau de contrôle sur le processus cessent d'être pertinents pour le calcul de la dimension implication⁴. Procéder autrement pour le codage de l'étendue de l'implication serait en effet contre-productif et entraînerait un double comptage des autres dimensions. Si les bénéficiaires d'un programme participent à deux étapes de l'évaluation, par exemple la conception de l'évaluation et la formulation des jugements et des recommandations, le score pour l'étendue de l'implication est de 0,5 (implication modérée). Ce score reste identique même si des décideurs participent aux mêmes tâches et si les parties prenantes impliquées exercent une influence plus importante sur ces tâches.

TABLEAU 13.1

Schème de codage pour l'étendue de l'implication

Étapes au sein desquelles les participants sont impliqués (nombre)	Niveau d'appartenance	
	Étiquette intuitive	Valeur numérique
0	Implication <i>nulle</i>	0
1	Implication <i>limitée</i>	0,25
2	Implication <i>modérée</i>	0,5
3	Implication <i>substantielle</i>	0,75
4	Implication <i>totale</i>	1

4. Le seuil minimal pour qu'il y ait implication est qu'au moins *un* type de partie prenante non évaluative participe à l'exécution d'une tâche en exerçant un niveau de contrôle faible/limité (voir la dimension contrôle).

TABLEAU 13.2**Typologie des parties prenantes non évaluatives**

Types	Description	Exemples tirés de l'extension du concept
Décideurs, concepteurs et gestionnaires du programme	Personnes politiquement, légalement et/ou organisationnellement responsables du programme et de son évaluation	Titulaire de charge publique (élu ou nommé), haut fonctionnaire, gestionnaire de mise en œuvre, président et chef de direction d'un organisme sans but lucratif, etc.
Responsables de la mise en œuvre du programme	Fonctionnaires et professionnels responsables de dispenser les biens et services du programme aux groupes cibles/bénéficiaires	Fonctionnaire de première ligne, infirmière, comptable, psychologue, agent de développement international, travailleur social, enseignant, agronome, etc.
Bénéficiaires directs et indirects; tiers négativement affectés par un programme	Personnes visées par le programme pour modifier leur comportement et/ou améliorer leur bien-être (groupes cibles et bénéficiaires); autres personnes affectées positivement ou négativement par le programme	Élève du niveau primaire, parent, gais et lesbiennes, famille immédiate d'un délinquant juvénile, conducteur en état d'ébriété, habitant d'un village nordique touché par l'exploitation minière, etc.
Société civile et citoyens	Groupes organisés et citoyens ayant un intérêt matériel, politique et/ou scientifique dans le programme et son évaluation	Groupe d'intérêts, syndicat, groupe de réflexion (<i>think tank</i>), expert indépendant, intellectuel, comité citoyen, etc.

Diversité des participants

Pour plus de simplicité, la deuxième dimension (*sélection des parties prenantes participant à l'évaluation*) a été rebaptisée *diversité des participants*. Nous considérons que plus les parties prenantes impliquées sont diverses, plus l'évaluation est participative, toutes choses étant égales par ailleurs. Comme précédemment, le terme *participants* désigne les parties prenantes non évaluatives.

La typologie présentée au tableau 13.2 englobe les catégories définies par Greene (2005) et s'étend au-delà. Les parties prenantes non évaluatives sont classées selon leur distance par rapport au programme et à la gestion de celui-ci : les décideurs en sont le plus près tandis que la société civile

en est le plus éloignée. S'il n'y a rien d'essentiel dans cette relation (par exemple, les syndicats et les fonctionnaires de première ligne pourraient intervenir directement dans la gestion du programme selon des cadres corporatistes), nous estimons qu'elle représente assez fidèlement la structure de programme pour la plupart des contextes. De plus, une typologie à quatre catégories est utile pour la mesure, car elle permet de distinguer facilement les différentes évaluations en fonction de leurs participants. La typologie est censée s'appliquer à toutes sortes de contextes et de finalités, par exemple l'évaluation démocratique de projets internationaux d'aide au développement. Selon la conceptualisation proposée, le fait qu'au moins deux organismes (par exemple, les ministères de l'Environnement et de la Santé) parrainent ensemble une évaluation ne suffit pas pour qualifier cette évaluation de participative. Les parties prenantes non évaluatives doivent contribuer directement au processus évaluatif.

TABLEAU 13.3**Schème de codage pour la diversité des participants**

Types de participants impliqués (nombre)	Niveau d'appartenance	
	Étiquette intuitive	Valeur numérique
0	Implication <i>nulle</i>	0
1	Implication <i>limitée</i>	0,25
2	Implication <i>modérée</i>	0,5
3	Implication <i>substantielle</i>	0,75
4	Implication <i>totale</i>	1

Le système de codage proposé considère chacune des catégories de parties prenantes comme un indicateur dichotomique (oui/non) de leur implication dans une évaluation donnée. Pour être jugée participative, une évaluation doit faire intervenir au moins un type de partie prenante non évaluative, de telle sorte que la pondération 0,25 soit attribuée à chacun des indicateurs constituant cette dimension secondaire (voir le tableau 13.3). Aucun type de parties prenantes non évaluatives (par exemple, responsables de la mise en œuvre du programme) n'est nécessaire pour considérer qu'une évaluation est participative. En outre, l'implication des divers participants n'est pas forcément progressive sur le plan des types. Ainsi, rien n'empêche des représentants de syndicats ou de groupes de pression (c'est-à-dire des groupes de la société civile) d'être les seules parties prenantes non évaluatives participant à une évaluation.

Nous avons évoqué la nécessité d'établir un cadre qui illustre bien l'aspect dynamique du processus évaluatif. Par exemple, le personnel de première ligne intervient peut-être dans la plupart des étapes du processus évaluatif, alors que les bénéficiaires d'un programme se contentent de définir les questions et les éléments d'évaluation. L'unité d'analyse pour la dimension *diversité* est l'évaluation comme telle, non les différentes étapes de son processus. Même si deux types de parties prenantes interviennent ensemble dans une étape du processus ou séparément dans deux tâches distinctes, la diversité des participants reste la même. Qui plus est, une évaluation faisant intervenir deux types de parties prenantes dans une étape du processus n'est pas fondamentalement différente d'une évaluation où l'intervention se poursuit tout au long du processus *du point de vue de la dimension diversité*. Par conséquent, cette dimension mesure la diversité des participants à une évaluation donnée et non la diversité à différentes étapes. Si, par exemple, le personnel de première ligne et les bénéficiaires directs d'un programme participent à une évaluation donnée (en ne tenant pas compte des tâches particulières dans lesquelles ils interviennent), cette évaluation en particulier obtiendrait une note de 0,5 pour la diversité, car deux types de parties prenantes non évaluatives sont présents.

Contrôle du processus évaluatif

L'évaluation participative se caractérise par le fait que les parties prenantes non évaluatives exercent un certain contrôle sur le processus évaluatif. L'hypothèse qui sous-tend cette dimension est que plus ce contrôle est important, plus l'évaluation est participative, toutes choses étant égales par ailleurs. Cette dimension est théorisée et mesurée de manière relative. Il faut donc comparer le contrôle qu'exercent les participants (dans l'ensemble) sur le processus à celui qu'exercent le ou les évaluateurs⁵.

5. Le cadre original (Daigneault et Jacob, 2009) considérait le contrôle comme distribué entre les participants, d'une part, et l'évaluateur ainsi que le commanditaire de l'évaluation, d'autre part. Or, puisque le commanditaire peut également être un participant *direct* à l'évaluation, nous avons opté pour une opérationnalisation du contrôle qui fait du commanditaire un participant (lorsque celui-ci l'est effectivement) ou ne le prend pas en compte (lorsque son rôle se limite à rédiger le cahier des charges et le contrat d'évaluation).

L'opérationnalisation de cette dimension est moins simple que celle des deux dimensions précédentes. D'une part, les indicateurs objectifs permettant de mesurer le contrôle (par exemple, spécification des responsabilités dans un contrat) laissent à désirer, car ils ne décrivent pas nécessairement comment le contrôle est partagé *en pratique* dans une évaluation particulière⁶. D'autre part, le contrôle peut varier de manière importante durant le processus (Themessl-Huber et Grutsch, 2003). À la différence de la diversité des participants, qui est mesurée pendant toute l'évaluation, le contrôle est indissociable des différentes tâches du processus. Par conséquent, le score doit correspondre à la manière dont le contrôle s'exerce à divers moments du processus, même si le contrôle est mesuré pendant toute l'évaluation.

TABEAU 13.4**Schème de codage pour le contrôle du processus évaluatif**

Niveau d'appartenance	
Étiquette intuitive	Valeur numérique
Contrôle <i>nul</i> des participants (contrôle total de l'évaluateur)	0
Contrôle <i>limité</i> des participants (contrôle substantiel de l'évaluateur)	0,25
Contrôle <i>égal</i> des participants et de l'évaluateur	0,5
Contrôle <i>substantiel</i> des participants (contrôle limité de l'évaluateur)	0,75
Contrôle <i>total</i> des participants (contrôle nul de l'évaluateur)	1

Nous proposons un schème de codage plus intuitif et souple que celui réservé aux autres dimensions de l'évaluation participative. Ce schème, qui repose sur un jugement nécessairement subjectif mais éclairé sur la répartition du pouvoir décisionnel, a recours à la même échelle de mesure que celle des autres dimensions à des fins d'agrégation (voir tableau 13.4). Une première étape consiste à se demander si le pouvoir décisionnel est réparti également entre l'évaluateur et les participants. Le cas échéant, le score de 0,5 est attribué pour le contrôle. Dans le cas contraire, il faut déterminer qui, des participants ou de l'évaluateur, a le plus d'influence.

6. La conceptualisation originale (Daigneault et Jacob, 2009) proposait deux indicateurs du contrôle : 1) l'autorité de prendre des décisions ; et 2) les autres ressources du contrôle. Comme ces indicateurs sont difficiles à appliquer, il a été décidé de les écarter.

Une fois que la direction de la balance du pouvoir est définie, il convient de déterminer s'il est totalement entre les mains des participants ou de l'évaluateur. Dans le premiers cas, les participants agissent à titre d'évaluateurs ou considèrent l'évaluateur comme un simple exécutant de leurs décisions. Dans le second cas, c'est l'évaluateur qui dirige exclusivement le processus. Les participants ont alors un simple rôle d'observateur ou de fournisseur de données. La dernière étape consiste à établir les variations qui peuvent survenir dans le niveau de contrôle au cours des différentes étapes du processus évaluatif. En cas de variation, il faut arriver à un score moyen ou du moins représentatif de l'ensemble du processus.

Combiner les dimensions pour mesurer le niveau de participation d'une évaluation

Les dimensions de l'étendue de l'implication, de la diversité des participants et du contrôle du processus sont les attributs fondamentaux, les dimensions constitutives de la participation. Quelle logique unit ces dimensions au concept général de participation ? Les dimensions du concept d'évaluation participative répondent selon nous à la *structure des conditions nécessaires et suffisantes* (voir Goertz, 2006). Cela signifie que, pour qu'une évaluation soit considérée comme participative, *chacune* des dimensions doit être présente. Par « présente », nous entendons un score de 0,25 ou plus sur cette dimension. Ces trois dimensions *suffisent* par ailleurs à définir le concept ; aucune autre dimension n'est requise (figure 13.1)⁷.

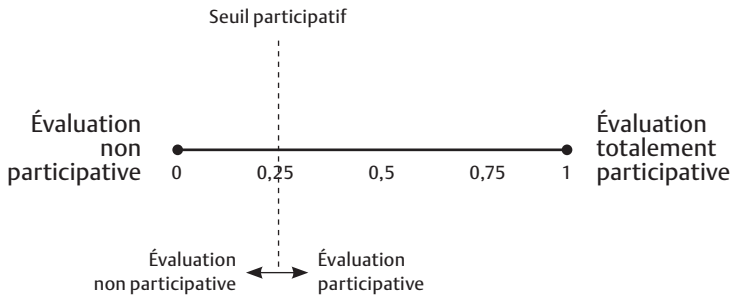
La logique d'agrégation des conditions nécessaires et suffisantes implique que le score du concept global, la participation en l'occurrence, est le *minimum* des dimensions (Goertz, 2006). Supposons qu'une évaluation englobe tous les types de parties prenantes non évaluatives et qu'elle produise de ce fait un score de 1 pour la dimension diversité. Cette évaluation obtiendrait un score global de 1 pour la participation *seulement si* les parties prenantes jouaient un rôle actif dans toutes les tâches d'évaluation (étendue de l'implication = 1) *et si* elles exerçaient un contrôle total sur la prise de décision pendant tout le processus en raison du pouvoir et des ressources officieuses qu'elles détiennent (contrôle = 1). Supposons plutôt que le score pour le contrôle est de 0,5, la même évaluation obtien-

7. Voir cependant la version révisée de l'instrument présentée en conclusion.

draît alors un score de 0,5 sur le plan du concept général. Si le score obtenu pour la dimension contrôle était nul, l'évaluation ne serait alors pas considérée comme participative, même si elle avait une note parfaite pour les deux autres dimensions⁸.

FIGURE 13.1

Indice de participation et deux concepts opposés



Un cas d'application : l'évaluation du collège des soins infirmiers de l'Université du Colorado à Denver

Cette conceptualisation de l'ÉP permet aux chercheurs et aux évaluateurs de mesurer le niveau de participation d'un processus d'évaluation. Elle permet aussi de comparer différents dispositifs d'évaluation, comme ce fut le cas du collège des soins infirmiers de l'Université du Colorado à Denver (Connors et Magilvy, 2011)⁹.

Ce collège emploie une centaine de professeurs qui, en plus d'être actifs en recherche et en développement de pratiques professionnelles, forment environ 800 étudiants aux différents cycles d'études. Dès sa fondation en 1998, le collège s'est doté d'un dispositif d'évaluation organisé autour d'un

8. La version originale de la conceptualisation proposait deux idéaux-types, « évaluation technocratique » et « évaluation démocratique autogérée », pour désigner les bornes du continuum conceptuel (Daigneault et Jacob, 2009). Ces termes ont toutefois été abandonnés en raison de leur charge normative négative et positive, respectivement. En outre, la logique qui unit les dimensions au concept a été révisée (voir la conclusion).

9. Il importe de préciser que la version de l'instrument de mesure utilisée par Connors et Magilvy (2011) est celle qui a été présentée dans l'*American Journal of Evaluation* (voir Daigneault et Jacob, 2009). Comme il a été mentionné à la note 1, l'instrument a été légèrement révisé depuis.

évaluateur interne (un professeur du collège) et d'un comité d'évaluation¹⁰ qui se réunit mensuellement pour réviser le plan d'évaluation du collège et examiner les résultats des évaluations effectuées par l'évaluateur interne. En 2006, à la suite d'un changement à la direction de la faculté et du départ à la retraite de l'évaluateur interne, le dispositif d'évaluation a été réformé. Les deux principaux changements sont l'engagement d'un évaluateur externe et la participation active du comité d'évaluation et de la nouvelle vice-doyenne aux évaluations.

La conceptualisation et l'instrument de mesure de la participation ont été appliqués aux deux systèmes d'évaluation par des parties prenantes à l'évaluation. Les deux évaluations sont considérées comme participatives puisqu'elles obtiennent respectivement un score final de 0,25 et de 0,50. Le tableau de la page suivante présente les résultats détaillés de chacune des évaluations.

Ce tableau révèle que l'augmentation du score global de la participation s'explique principalement par une plus grande implication du comité d'évaluation dans toutes les étapes du processus et un contrôle accru sur les décisions et les actions liées à la réalisation de l'évaluation. Le score relatif à la diversité des participants reste inchangé.

En utilisant l'indice de participation, certains pourraient penser, à tort, qu'il est empreint de normativité et qu'il vise à accroître la participation. Or, comme nous l'avons vu dans ce chapitre, la participation n'est pas une fin en soi. Au contraire, l'interprétation et la discussion autour du score global d'une évaluation permettent d'entamer une réflexion pour savoir s'il est nécessaire, souhaitable et faisable d'accroître ce score. Dans le cas du collège des soins infirmiers, les auteures reconnaissent qu'en raison de la nature du fonctionnement de leurs instances universitaires, il n'est pas possible d'étendre la participation aux bénéficiaires du programme (les étudiants) ou aux citoyens. Par conséquent, même si les auteures pensent que l'ajout de leur point de vue serait souhaitable, le score global de la participation ne pourra pas dépasser 0,5 même si le score est augmenté sur une autre dimension. En conclusion, Connors et Magilvy (2011) affirment que l'indice est « facile à comprendre, à utiliser et qu'il est pertinent » (p. 82 : traduction). Elles estiment également que les « scores étaient

10. Ce comité est composé de professeurs, des directeurs de programme du collège des soins infirmiers et d'administrateurs universitaires.

conformes à leur intuition selon laquelle la participation s'était accrue en raison de la réforme et de l'implication d'un évaluateur externe » (p. 83 : traduction).

Conclusion : un instrument encore plus en phase avec les intuitions des évaluateurs

La conceptualisation et l'opérationnalisation de l'ÉP qui ont été décrites dans ce chapitre constituent un important progrès dont peuvent bénéficier les chercheurs et les praticiens qui s'intéressent à l'évaluation. Les chercheurs, en particulier, peuvent désormais compter sur un instrument de mesure dont les résultats de validation quantitatifs (Daigneault, Jacob et Tremblay, en préparation) et mixtes (c'est-à-dire combinant méthodes quantitatives et qualitatives) (Daigneault et Jacob, à paraître) confirment de manière générale la fidélité et la validité des scores produits par l'instrument de mesure (appelé *Participatory Evaluation Measurement Instrument*).

TABEAU 13.5

Niveau de participation de l'évaluation du collège des soins infirmiers de l'Université du Colorado à Denver

	Valeur	Étiquette	Justification
Évaluation initiale			
Étendue de l'implication	0,25	Implication limitée	Les tâches des membres du comité d'évaluation se situent sur le plan de la conception de l'évaluation (révision du plan d'évaluation et avis sur le processus évaluatif).
Diversité des participants	0,5	Diversité modérée	Le comité d'évaluation réunit deux groupes de parties prenantes non évaluatives : les décideurs (vice-doyenne et directeurs de programme) et les responsables de la mise en œuvre du programme (professeurs et personnel enseignant).
Contrôle du processus évaluatif	0,25	Contrôle limité des participants	L'évaluateur interne exerce un contrôle substantiel sur l'évaluation en maîtrisant l'expertise technique.
Score global	0,25	Évaluation faiblement participative	

	Valeur	Étiquette	Justification
Évaluation après la réforme			
Étendue de l'implication	1	Implication totale	Les parties prenantes non évaluatives participent aux quatre étapes du processus d'évaluation. Le comité d'évaluation continue de réviser le plan d'évaluation et de fournir des avis sur le processus d'évaluation. À cela s'ajoute une participation dans le développement d'instruments de collecte de données, l'interprétation des résultats, la formulation d'un jugement, la présentation et la diffusion des résultats auprès d'autres décideurs du collège des soins infirmiers.
Diversité des participants	0,50	Diversité modérée	Le comité d'évaluation réunit deux groupes de parties prenantes non évaluatives : les décideurs (vice-doyenne et directeurs de programme) et les responsables de la mise en œuvre du programme (professeurs et personnel enseignant). À la différence de l'évaluation initiale, l'évaluation après la réforme associe à ses travaux d'autres professeurs que ceux qui siègent sur le comité d'évaluation. Étant donné qu'il s'agit de membres appartenant à la même catégorie que celle déjà présente, le score sur cette dimension n'augmente pas malgré ce changement.
Contrôle du processus évaluatif	0,75	Contrôle substantiel des participants	Le comité d'évaluation contrôle davantage le processus évaluatif depuis qu'il travaille avec un évaluateur externe.
Score global	0,50	Évaluation modérément participative	

Les résultats de l'étude de validation mixte nous ont toutefois amenés à réviser la conceptualisation et l'instrument de mesure. Plusieurs utilisateurs de l'instrument ont en effet souligné que le recours au minimum des dimensions pour déterminer le score global de participation était contre-intuitif, car trop conservateur (un score élevé est difficile à atteindre). Nous avons défendu la thèse selon laquelle les dimensions d'étendue de l'implication, de diversité des participants et de contrôle sont nécessaires au concept de participation. Or, dans ce cas, le minimum est incontournable. Il existe toutefois une structure conceptuelle hybride qui tient compte à la fois de nos intuitions et de celles des évaluateurs.

Une version révisée de l'instrument s'appuyant sur une structure hybride a donc été proposée (voir Daigneault et Jacob, à paraître). Cette version révisée s'articule en deux temps. Tout d'abord, les dimensions de diversité des participants, d'étendue de l'implication et de contrôle du processus évaluatif demeurent nécessaires pour distinguer les cas participatifs des cas non participatifs. Une évaluation entièrement contrôlée par l'évaluateur est ainsi considérée comme non participative. Une fois passé ce filtre, le score global est déterminé par la *moyenne* des dimensions¹¹. Ainsi, avec des scores de 0,5, 1 et 0,75, l'évaluation postréforme de Connors et Magilvy (2011) se verrait attribuer un score global de participation de 0,75 avec la version révisée, au lieu de 0,5 avec la version originale. La révision amène en fait les scores élevés à « compenser » en partie les scores faibles.

Dans le cadre de l'étude de validation mixte, une majorité de répondants dont les scores étaient modifiés par la version révisée ont affirmé préférer celle-ci à l'ancienne (Daigneault et Jacob, à paraître). Ce résultat vient appuyer la révision proposée et nous ne pouvons que recommander d'utiliser cette version dans le cadre de recherches ultérieures.

Références

- BURKE, B. (1998), « Evaluating for a change: Reflections on participatory methodology », *New Directions for Evaluation*, 80, 43-56.
- BUTTERFOSS, F. D., V. FRANCISCO et E. M. CAPWELL (2001), « Stakeholder participation in evaluation », *Health Promotion Practice*, 2(2), 114-119.
- CHRISTIE, C. A. (2003), « What guides evaluation? A study of how evaluation practice maps onto evaluation theory », *New Directions for Evaluation*, 97, 7-35.
- CONNORS, S. C. et J. K. MAGILVY (2011), « Assessing vital signs: Applying two participatory evaluation frameworks to the evaluation of a college of nursing », *Evaluation and Program Planning*, 34(2), 79-86.
- COUSINS, J. B. (2001), Do evaluator and program practitioner perspective converge in collaborative evaluation? *Revue canadienne d'évaluation de programmes*, 16(2), 113-133.
- COUSINS, J. B., J. J. DONOHUE et G. A. BLOOM (1996), « Collaborative evaluation in North America: Evaluators' self-reported opinions, practices, and consequences », *American Journal of Evaluation*, 17(3), 207-226.

11. Au besoin, cette moyenne est arrondie au score inférieur.

- COUSINS, J. B. et L. M. EARL (1992), « The case for participatory evaluation », *Educational Evaluation and Policy Analysis*, 14(4), 397-418.
- COUSINS, J. B. et L. M. EARL (1995), « Participatory evaluation in education : What do we know ? Where do we go ? », dans COUSINS, J. B. et L. M. EARL (dir.), *Participatory Evaluation in Education : Studies in Evaluation Use and Organizational Learning* (p. 159-180), Londres, Falmer Press.
- COUSINS, J. B. et E. WHITMORE (1998), « Framing participatory evaluation », *New Directions for Evaluation, Understanding and Practicing Participatory Evaluation*, 80, 5-23.
- DAIGNEAULT, P.-M. et S. JACOB (2009), « Toward accurate measurement of participation : Rethinking the conceptualization and operationalization of participatory evaluation », *American Journal of Evaluation*, 30(3), 330-348.
- DAIGNEAULT, P.-M. et S. JACOB (2012), « Les concepts souffrent-ils de négligence bénigne en sciences sociales ? Éléments d'analyse conceptuelle et examen exploratoire de la littérature francophone à caractère méthodologique », *Information sur les sciences sociales*, 51(2), 188-204.
- DAIGNEAULT, P.-M. et S. JACOB (à paraître), « Unexpected but most welcome : Mixed methods or the validation and revision of the participatory evaluation measurement instrument », *Journal of Mixed Methods Research*.
- DAIGNEAULT, P.-M., S. JACOB et J. TREMBLAY (à paraître), « Measuring stakeholder participation in evaluation : An empirical validation of the participatory evaluation measurement instrument (PEMI) », *Evaluation review*.
- FLEISCHER, D. N. et C. A. CHRISTIE (2009), « Evaluation use results from a survey of US American Evaluation Association members », *American Journal of Evaluation*, 30(2), 158-175.
- GERRING, J. (1999), « What makes a concept good ? A criterial framework for understanding concept formation in the social sciences », *Polity*, 31(3), 357-393.
- GOERTZ, G. (2006), *Social Science Concepts : A User's Guide*, Princeton, Princeton University Press.
- GREENE, J. C. (1987), « Stakeholder participation in evaluation design : Is it worth the effort ? », *Evaluation and Program Planning*, 10(4), 379-394.
- GREENE, J. C. (2005), « Stakeholders », dans MATHISON, S. (dir.), *Encyclopedia of Evaluation* (p. 397-398), Thousand Oaks, Sage.
- GREGORY, A. (2000), « Problematizing participation : A critical review of approaches to participation in evaluation theory », *Evaluation*, 6(2), 179-199.
- HUBERMAN, M. (1995), « The many modes of participatory evaluation », dans COUSINS, J. B. et L. M. EARL (dir.), *Participatory Evaluation in Education : Studies in Evaluation Use and Organizational Learning* (p. 103-111), Londres, Falmer Press.

- JACOB, S. et P.-M. DAIGNEAULT (2011), « La gouvernance et l'implication des parties prenantes dans l'évaluation des politiques: panacée ou boîte de pandore? », dans ROUILLARD, C. et N. BURLONE (dir.), *L'État et la société civile sous le joug de la gouvernance* (p. 217-242), Québec, Presses de l'Université Laval.
- JACOB, S. et L. OUVREAU (2009), « L'évaluation participative. Avantages et difficultés d'une pratique innovante », *Cahiers de la performance et de l'évaluation*, 1, 82: www.pol.ulaval.ca/Perfeval/upload/2009__cahiers_perfeval_1_participation._04112009_93426.pdf
- KING, J. A. (1998), « Making sense of participatory evaluation practice », *New Directions for Evaluation*, 80, 57-67.
- KING, J. A. (2005), « Participatory evaluation », dans Mathison, S. (dir.), *Encyclopedia of Evaluation* (p. 291-294), Thousand Oaks, Sage.
- MARK, M. M. (2001), « Evaluation's future: Furor, futile, or fertile? », *American Journal of Evaluation*, 22(3), 457-479.
- MATHIE, A. et J. C. GREENE (1997), « Stakeholder participation in evaluation: How important is diversity? », *Evaluation and Program Planning*, 20(3), 279-285.
- MATHISON, S. (2005), « Preface », dans MATHISON, S. (dir.), *Encyclopedia of Evaluation* (p. 33-35), Thousand Oaks, Sage.
- MURRAY, R. (2002), « Citizens' control of evaluations: Formulating and assessing alternatives », *Evaluation*, 8(1), 81-100.
- REBIEN, C. C. (1996), « Participatory evaluation of development assistance: Dealing with power and facilitative learning », *Evaluation*, 2(2), 151-171.
- RIDDE, V. (2006), « Suggestions d'améliorations d'un cadre conceptuel de l'évaluation participative », *Revue canadienne d'évaluation de programmes*, 21(2), 1-23.
- ROSSI, P. H., M. W. LIPSEY et H. E. FREEMAN (2004), *Evaluation: A Systematic Approach* (7^e éd.), Thousand Oaks, Sage.
- SARTORI, G. ([1984] 2009), « Guidelines for concept analysis », dans D. COLLIER et J. GERRING (dir.), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori* (p. 97-150), Londres, Routledge.
- SCHEDLER, A. (2011), « Concept formation », dans BADIE, B., D. BERG-SCHLOSSER et L. A. MORLINO (dir.), *International Encyclopedia of Political Science* (p. 370-382), Thousand Oaks, Sage.
- SCRIVEN, M. (1997), « Empowerment evaluation examined », *American Journal of Evaluation*, 18(1), 165-175.
- SCRIVEN, M. (2005), « Book Review: Empowerment evaluation principles in practice », *American Journal of Evaluation*, 26(3), 415-417.
- SHEA, M. P. et J. H. LEWKO (1995), « Use of a stakeholder advisory group to facilitate the utilization of evaluation results », *Revue canadienne d'évaluation de programmes*, 10(1), 159-162.

- THEMESSL-HUBER, M. T. et M. A. GRUTSCH (2003), « The shifting locus of control in participatory evaluations », *Evaluation*, 9(1), 92-111.
- WEAVER, L. et J. B. COUSINS (2004), « Unpacking the participatory process », *Journal of MultiDisciplinary Evaluation*, 1, 19-40.
- WEISS, C. H. (1986), « The stakeholder approach to evaluation: Origins and promise », dans HOUSE, E. R. (dir.), *New Directions in Educational Evaluation* (p. 145-157), Londres, Falmer Press.
- WHITMORE, E. (1998), « Editor's notes », *New Directions for Evaluation*, 80, 1-3.

14

Théorie et pratique de l'approche *Realist* pour l'évaluation des programmes*

*Valéry Ridde, Émilie Robert, Anne Guichard,
Pierre Blaise et Josefien Van Olmen*

Évoquer l'évaluation des interventions complexes ne peut assurément faire l'impasse d'un courant théorique relativement récent qui propose, parmi d'autres, une solution à cette difficulté évaluative. L'approche *Realist* est en effet présentée comme une approche d'évaluation permettant de produire des « données probantes sur les interventions sociales complexes, qui fournissent une explication de la manière dont celles-ci fonctionnent (ou non), et pour quelles raisons, dans des contextes spécifiques¹ » (Pawson, Greenhalgh, Harvey et Walshe, 2004). La perspective épistémologique est celle du réalisme critique, que nous expliciterons plus loin, différente du constructivisme et du post-positivisme. Nous avons délibérément choisi de ne pas traduire le terme « *Realistic* » ou « *Realist* » en

* Ce chapitre est d'abord paru sous la forme d'un article publié dans un numéro spécial de la *Revue canadienne d'évaluation de programmes*, 2012, 26 (3). L'article a été écrit avant mais publié après cet excellent texte auquel nous renvoyons le lecteur : Marchal *et al.* (2012), « Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research », *Evaluation*, 18(2).

1. Toutes les citations en anglais ont été traduites par les auteurs.

français, car la traduction littérale (« réaliste ») ne nous semble pas rendre compte du sens du terme. Son examen pourrait nous entraîner vers des débats concernant des enjeux liés notamment à la faisabilité des évaluations (par exemple, *Real World Evaluation*; *Evaluability Assessment*), ce qui n'est pas le sujet de ce texte. En effet, l'objectif ici est de présenter les tenants et les aboutissants de cette approche en mettant particulièrement en relief les défis méthodologiques de son emploi dans les évaluations (*Realistic Evaluation*) et les synthèses de connaissances (*Realist Review/Synthesis*). Pour ce faire, nous avons recours, d'une part, à l'étude des écrits scientifiques d'auteurs ayant fait appel à cette approche, d'autre part, à l'analyse réflexive de nos propres expériences au Canada, en France et en Afrique.

L'appréhension de la complexité par l'approche *Realist*

On ne reviendra pas ici sur les débats concernant la complexité traités par d'autres (Morin, 2005; Ridde, Delormier et Gaudreau, 2007). D'une manière globale, il faut concevoir les interventions complexes comme des enchevêtrements d'éléments multiples, unis la plupart du temps par des relations non linéaires, où la compréhension de la contribution d'un seul n'est pas suffisante pour analyser l'intervention comprise comme un tout. Ces interventions ont une histoire, elles s'inscrivent dans un contexte particulier qu'il faut prendre en compte. Elles sont des systèmes ouverts dont le chercheur qui les observe fait lui-même partie (Blaise, Marchal, Lefèvre et Kegels, 2010). Pour Pawson et ses collaborateurs (2004), les interventions complexes ont sept caractéristiques. Nous donnons, dans le tableau 14.1, un exemple de la nature de cette complexité à travers le cas des politiques qui visent à supprimer le paiement direct des soins récemment mises en place en Afrique. Compte tenu de cette complexité, le recours à des démarches évaluatives expérimentales (ou à des synthèses quantitatives de type méta-analyse) paraît peu pertinent (Blaise *et al.*, 2010), comme le montre bien une étude sur la suppression du paiement des soins en Afrique (Lagarde et Palmer, 2006).

S'inspirant des méthodes d'évaluation fondées sur la théorie (Astbury et Leeuw, 2010; Chen, 2005; Donaldson, 2007) et d'une épistémologie réaliste (Julnes, Mark et Henry, 1998), Pawson propose un cadre structurant pour décrire les logiques des interventions observées. L'approche

TABEAU 14.1**Les sept caractéristiques des interventions complexes appliquées aux politiques d'abolition du paiement des soins en Afrique**

Les interventions sociales...	Les politiques d'abolition du paiement direct...
<i>... sont des théories ou sont composées de plusieurs théories.</i>	<i>... visent à réduire le fardeau financier des ménages, à améliorer l'accès aux services de santé et leur recours précoce aux soins, etc.</i>
<i>... impliquent la participation de nombreuses parties prenantes.</i>	<i>... impliquent les États, la communauté internationale, les ONG, les populations, les agents de santé, etc.</i>
<i>... consistent en une chaîne de processus décisionnels.</i>	<i>... supposent l'identification du problème, la formulation de la politique, la mise en œuvre de la politique par différents acteurs, l'acceptation de la politique par la population, etc.</i>
<i>... ne sont pas linéaires et sont soumises à des boucles de rétroaction.</i>	<i>... se transforment et s'adaptent par l'action et l'influence des agents de santé, des patients, des décideurs, etc.</i>
<i>... sont enchâssées dans plusieurs systèmes sociaux et plusieurs contextes.</i>	<i>... sont mises en œuvre dans plusieurs pays, eux-mêmes composés de différentes populations, elles-mêmes évoluant dans des réalités sociales et des représentations du monde distinctes, et ayant des systèmes de santé différents.</i>
<i>... sont perméables à l'influence d'autres interventions.</i>	<i>... coexistent avec d'autres politiques de santé ou interventions sociales qui ont une influence sur elles.</i>
<i>... sont des systèmes ouverts apprenants.</i>	<i>... sont des systèmes dans lesquels les agents de santé, les patients, les décideurs, etc., adaptent leurs pratiques, leurs comportements, leurs attitudes, etc.</i>

Realist dispose souvent d'un caractère exploratoire et permet de mettre au jour, de tester et d'améliorer des théories dites de moyenne portée. Selon Merton (1968), il s'agit de « théories dont le niveau se situe entre [d'abord] les hypothèses de travail, mineures mais nécessaires, qui se développent au quotidien dans le cadre de la recherche, et [ensuite] les efforts systématiques d'intégration pour développer une théorie unitaire qui expliquerait toutes les uniformités observées du comportement social, de l'organisation sociale, et du changement social ». Ces théories concernent des mécanismes de changement qui sont liés aux raisonnements humains ainsi que les choix des personnes et leurs aptitudes (Pawson et Tilley, 1997). Car, si les programmes fonctionnent, c'est bien parce que les personnes en ont décidé ainsi. Il s'agit donc de mettre en lumière le mécanisme (M) actionné par l'intervention pour produire ses effets (E) dans un contexte donné (C), ce que Pawson appelle la configuration « contexte-mécanisme-effet » (CME). Il suggère ensuite de comparer la CME à la

« théorie² de l'intervention », autrement dit à la logique de l'intervention qui décrit la chaîne des résultats et la production des effets. Le tableau 14.2 présente quelques définitions des concepts clés de l'approche *Realist*.

TABEAU 14.2
Les concepts clés de l'approche *Realist*

Mécanisme	Élément du raisonnement de l'acteur face à une intervention. Un mécanisme : 1) est généralement caché ; 2) est sensible aux variations du contexte ; 3) produit des effets.
Configuration C-M-E	Outil conceptuel permettant de mettre en relation les éléments du contexte, les mécanismes et les effets d'une intervention.
Théorie de l'intervention	Ensemble des hypothèses qui expliquent la manière dont (comment et pourquoi) l'intervention est censée produire ses effets. Elle peut se décliner sous la forme d'une ou plusieurs C-M-E.
Théorie de moyenne portée	Niveau d'abstraction théorique qui permet d'expliquer les tendances et les régularités observées dans les interactions contexte-mécanisme-effet d'un ensemble d'interventions.

Adaptés de Astbury et Leeuw, 2010 ; Blaise *et al.*, 2010 ; Pawson et Tilley, 1997.

Cette démarche s'éloigne des approches expérimentales qui cherchent à tester la stabilité d'une relation causale entre une intervention et son effet escompté, et qui considèrent les circonstances et les détails contextuels uniquement comme des variables numériques et des facteurs de confusion. Les tenants de l'approche *Realist* intègrent le contexte dans la production d'effets. La recherche ne vise pas à obtenir une réponse définitive sur l'intervention la plus efficace dans toutes les circonstances (Blaise *et al.*, 2010). L'objectif n'est plus, par exemple, de répondre à la question « l'abolition du paiement fonctionne-t-elle ? », mais « est-ce que les politiques d'abolition du paiement en Afrique fonctionnent, comment, pour qui et dans quelles circonstances particulières ? » (Robert, Ridde, Marchal et Fournier, 2012).

Rendre compte de la manière dont fonctionne la logique d'intervention (et les CME) dans un contexte spécifique avec un cadre d'analyse *Realist*

2. Nous mettons ici cette expression entre guillemets, car le terme « théorie » est souvent galvaudé dans le domaine de l'évaluation de programmes (parlons de logique d'intervention plutôt), comme le reconnaît par exemple Alkin (2004), ce qui ne l'empêche pas d'appeler « *theorists* » les tenants de certaines *approches* évaluatives (terme que nous préférons).

permet de projeter les résultats de recherche dans différents contextes. Partant du postulat qu'il est possible d'observer différentes interventions suivant la même logique — une même famille de mécanismes — dans des contextes différents, l'approche *Realist* suggère la possibilité de mettre au jour des régularités (*patterns*) liées à des caractéristiques spécifiques du contexte dans les effets constatés. En réalisant de telles observations empiriques (ou des synthèses) multiples dans des contextes variés, il est *a priori* possible de voir apparaître ces régularités. On cherche non pas à vérifier des effets, constants et reproductibles, en fonction de variables ou de facteurs de confusion, mais plutôt à observer des régularités ou des tendances dans les effets produits (positifs, négatifs ou inattendus), dans des contextes semblables ou différents, mais dont les différences sont mises au jour, puis caractérisées dans une typologie.

Le fait de trouver des effets variés n'est pas en soi un problème. Ce que l'on veut comprendre, c'est pourquoi un mécanisme produit des effets différents dans un autre contexte, ou bien provoque des effets identiques dans des contextes différents ou des effets différents dans des contextes identiques. En repérant les régularités qui surviennent dans ces relations, il serait possible de les regrouper dans une typologie de CME, et ainsi d'améliorer la théorie explicative des régularités observées (Pawson, 2003). De CME en CME, on améliore graduellement notre compréhension des mécanismes qui entrent en jeu dans une intervention (Pawson, 2002). Il s'agit ici d'appliquer la démarche analytique proposée par Yin (1994) pour la réalisation d'une étude de cas qui consiste à construire une « explication de nature itérative » : « les données issues de l'étude de cas sont examinées, les positions théoriques sont amendées puis les données sont une nouvelle fois examinées à partir d'une nouvelle perspective, selon cette démarche itérative » (p. 111).

L'approche *Realist* peut être utilisée dans le cadre de synthèses, ou revues de la littérature, et dans le cadre d'évaluations. La synthèse *Realist* fait partie des méthodes qui servent à résumer des connaissances déjà disponibles, à l'instar des synthèses narratives ou des méta-analyses. Elle s'inscrit dans le récent courant des *Mixed Studies Reviews* (Grant et Booth, 2009 ; Pope, Mays et Popay, 2007). Elle prend en compte tous les types d'études, qu'elles soient de nature quantitative, qualitative ou mixte (Pluye *et al.*, 2009), et elle permet de considérer tous les écrits qui apportent des informations supplémentaires sur l'intervention étudiée. Nous considérons que la synthèse *Realist* est une démarche d'évaluation, au même titre

que l'étude de cas ou l'étude observationnelle, dans la mesure où elle permet de comprendre et/ou de porter un jugement sur une intervention (Ridde et Dagenais, 2009). L'évaluation *Realist* implique une collecte de données primaires au cours de l'évaluation d'une intervention. Nous allons maintenant évoquer les défis particuliers de l'utilisation de l'approche *Realist* dans le cadre de ces deux démarches.

Les utilisations de l'approche *Realist* rapportées dans les écrits scientifiques

L'opérationnalisation de la démarche

L'approche *Realist* étant récente, peu de chercheurs se sont encore essayés à son exercice. En santé publique, dix synthèses, neuf études empiriques et une recherche combinant synthèse avec évaluation *Realist* ont été recensées (tableau 14.3). Une recherche documentaire dans les bases de

TABLEAU 14.3
Articles rapportant une utilisation de l'approche *Realist* (1997-2010)

Synthèses
Clamp, S. et J. Keen (2007), « Electronic health records : Is the evidence base any use ? », <i>Informatics for Health and Social Care</i> , 32(1), 5-10.
Dieleman, M., B. Gerretsen et G. J. van der Wilt (2009), « Human resource management interventions to improve health workers' performance in low and middle income countries : A realist review », <i>Health Research Policy and System</i> , 7(7).
Greenhalgh, T., E. Kristjansson et V. Robinson (2007), « Realist review to understand the efficacy of school feeding programmes », <i>British Medical Journal</i> , 335, 858-861.
Jackson, L. et al. (2009), « Does moving from a high-poverty to lower-poverty neighborhood improve mental health ? A realist review of "Moving to Opportunity" », <i>Health et Place</i> , 15(4), 961-970.
Kane, S. S. et al. (2010), « A realist synthesis of randomised control trials involving use of community health workers for delivering child health interventions in low and middle income countries », <i>BMC Health Services Research</i> , 10(286).
Law, J., C. Plunkett, J. Taylor et M. Gunning (2009), « Developing policy in the provision of parenting programmes : Integrating a review of reviews with the perspectives of both parents and professionals », <i>Child : Care, Health and Development</i> , 35(3), 302-312.
Leeman, J. et al. (2010), « Implementation of antiretroviral therapy adherence interventions : A realist synthesis of evidence », <i>Journal of Advanced Nursing</i> , 66(9), 1915-1930.
McCormack, B. et al. (2007), « A realist synthesis of evidence relating to practice development : Methodology and methods », <i>Practice Development in Health Care</i> , 6(1), 5-24.
O'Campo, P. et al. (2009), « Community-based services for homeless adults experiencing concurrent mental health and substance use disorders : A realist approach to synthesizing evidence », <i>Journal of Urban Health : Bulletin of the New York Academy of Medicine</i> , 86(6), 965-989.
Wong, G., T. Greenhalgh et R. Pawson (2010), « Internet-based medical education : A realist review of what works, for whom and in what circumstances », <i>BMC Medical Education</i> , 10(1).

Études empiriques

- Byng, R., I. Norman et S. Redfern (2005), « Using realistic evaluation to evaluate a practice-level intervention to improve primary healthcare for patients with long-term mental illness », *Evaluation: The International Journal of Theory, Research and Practice*, 11(1), 69-93.
- Byng, R., I. Norman, S. Redfern et R. Jones (2008), « Exposing the key functions of a complex intervention for shared care in mental health: Case study of a process evaluation », *BMC Health Services Research*, 8, 274.
- Clark, A., H. K. Whelan, R. Barbour et P. D. MacIntyre (2005), « A realist study of the mechanisms of cardiac rehabilitation », *Journal of Advanced Nursing*, 52(4), 362-371.
- Evans, D. et A. Killoran (2000), « Tackling health inequalities through partnership working: Learning from a realistic evaluation », *Critical Public Health*, 10(2), 125-140.
- Kennedy, A., A. Rogers et C. Gately (2005), « Assessing the introduction of the expert patients programme into the NHS: A realistic evaluation of recruitment to a national lay-led self-care initiative », *Primary Health Care Research and Development*, 6(2), 137-148.
- Marchal, B., M. Dedzo et G. Kegels (2010), « A realist evaluation of the management of a well-performing regional hospital in Ghana », *BMC Health Services Research*, 10(24).
- Oroviogioicochea, C. et R. Watson (2009), « A quantitative analysis of the impact of a computerised information system on nurses' clinical practice using a realistic evaluation framework », *International Journal of Medical Informatics*, 78(12), 839-849.
- Tolson, D., J. McIntosh, L. Loftus et P. Cormie (2007), « Developing a managed clinical network in palliative care: A realistic evaluation », *International Journal of Nursing Studies*, 44(2), 183-195.
- Wilson, V., B. McCormack et G. Ives (2008), « Developing healthcare practice through action learning: Individual and group journeys », *Action Learning: Research and Practice*, 5(1), 21-38.

Étude combinant synthèse avec étude empirique

- Rycroft-Malone, J., M. Fontenla, D. Bick et K. Seers (2010), « A realistic evaluation: The case of protocol-based care », *Implementation Science*, 5(38).
-

données MEDLINE, CINAHL, EMBASE, CAB HEALTH, SAGE Journals Online et ERIC à l'aide des expressions clés suivantes : « *realist* review* », « *realist* synthesis* », « *realist* evaluation* » et « *realist* approach* » a permis de recenser les écrits publiés jusqu'en 2010. Malgré la nouveauté et la complexité de cette approche, certains articles ne font pas état des difficultés rencontrées (Clark, Whelan, Barbour et MacIntyre, 2005 ; Evans et Killoran, 2000 ; Greenhalgh, Kristjansson et Robinson, 2007 ; Jackson *et al.*, 2009 ; Kennedy, Rogers et Gately, 2005 ; Law, Plunkett, Taylor et Gunning, 2009 ; Oroviogioicochea et Watson, 2009 ; Wilson, McCormack et Ives, 2008). Pourtant, les auteurs appréhendent et appliquent différemment l'approche *Realist*, ce qui laisse deviner le défi majeur auquel on fait face : l'application d'une approche novatrice, récente et encore peu utilisée, pour laquelle Pawson a préféré donner des pistes plutôt qu'imposer un cadre méthodologique strict.

En effet, les ouvrages traitant de l'évaluation réaliste en expliquent les principes fondamentaux et définissent les éléments de l'équation CME, sans expliciter véritablement sa mise en pratique. Si cela permet une certaine latitude, comme l'indiquent Rycroft-Malone et ses collègues (2010), le processus de recherche peut devenir d'autant plus long que les chercheurs ont une expertise limitée dans le domaine. Les dimensions temporelle et financière associées à la démarche sont également à prendre en considération. La recherche menée par Marchal *et al.* (2010) montre bien la nécessité de s'interroger sur la faisabilité d'une évaluation de type réaliste. Dans leur cas, trois visites de terrain impliquant une collecte de données ont été réalisées entre 2004 et 2007, sans compter le temps consacré à l'analyse des données. Rycroft-Malone et ses collègues (2010) relèvent le même défi en précisant que, faute de ressources, ils n'ont pu mener entièrement une synthèse *Realist* ni tester autant de configurations CME que cela aurait été nécessaire.

Distinguer ce qui appartient au contexte, à l'intervention et au mécanisme

Au cours de la structuration des configurations CME, il est souvent difficile de déterminer si on est en présence d'un mécanisme, d'un contexte, voire d'une configuration CME, ce que Marchal et ses collègues (2010) appellent le « dilemme CME ». Ce dilemme s'est posé à plusieurs auteurs (Byng, Norman, Redfern et Jones, 2008). En effet, les chercheurs doivent arriver à centrer leur analyse au niveau où ils pourront observer des régularités CME. De la même façon que le photographe règle sa focale pour circonscrire son point d'intérêt, le chercheur doit adapter son angle d'analyse de manière à percevoir l'ensemble des mécanismes de l'intervention, identifier les régularités observées dans les contextes et les effets et développer des configurations CME. La confusion entre contexte, mécanisme et intervention surgit lorsque le focus du chercheur n'est pas réglé, c'est-à-dire lorsque le chercheur n'a pas encore tout à fait déterminé son niveau d'analyse. Ainsi il peut se trouver face à un mécanisme, un contexte ou un effet, selon le réglage de sa focale. Ce défi est bien relevé par McCormack et ses collègues dans leur article méthodologique (2007). Toutefois, cette confusion est surtout d'ordre pragmatique, car elle n'empêche pas de comprendre le cheminement causal entre les éléments CME dans le cadre d'une intervention. Comme Byng et ses collègues (2005) le

rappellent : « Dans la mesure où la notation CME a été vue comme une technique méthodologique, avec une distinction plus pragmatique que théorique entre contexte et mécanisme, il y a eu moins de problèmes. L'important est que la recherche des configurations CME a amélioré notre compréhension de ce qui s'était produit. » Parvenir à cette distinction afin d'élaborer les configurations CME est une entreprise de longue haleine.

Des défis inhérents à la synthèse des connaissances

L'identification des théories d'intervention est une gageure relevée par Dieleman et ses collègues (2009). En effet, ces théories sont rarement explicitées dans les travaux publiés, notamment du fait des contraintes de mots imposées par les revues. Ces contraintes expliquent également la difficulté pour les auteurs de synthèses de trouver des éléments contextuels (Kane *et al.*, 2010 ; O'Campo *et al.*, 2009) ou les mécanismes en œuvre (Dieleman *et al.*, 2009 ; Leeman *et al.*, 2010). Il peut aussi arriver que plusieurs théories soient sous-jacentes à une seule intervention, mais que les analyses portant sur cette intervention ne concernent qu'une ou certaines de ces théories. Les auteurs des revues *Realist*, comme Dieleman *et al.* (2009), Clamp et Keen (2007), et Wong *et al.* (2010), se trouvent alors dans l'impossibilité de tester et de comparer les différentes théories d'intervention, ce qui limite la portée de leur synthèse.

De plus, il peut s'avérer compliqué, voire impossible, de mettre au jour des régularités dans les configurations CME, et en particulier l'influence du contexte sur les effets. En effet, cela suppose, d'une part, de pouvoir recueillir suffisamment d'information au sujet du contexte, et d'autre part, de trouver des interventions similaires dans des contextes différents, ce que résumait bien Dieleman *et al.* (2009) : « Il n'a pas été possible de repérer des récurrences (*patterns*) sur la manière dont le contexte influe sur les effets des interventions étant donné leur description limitée et le fait que très peu d'interventions similaires sont mises en œuvre dans des contextes différents. »

Notre pratique de l'approche *Realist* fait écho aux défis évoqués dans la littérature. Nous proposons maintenant de les présenter plus en profondeur, en nous appuyant sur nos expériences en tant que chercheurs en évaluation.

Nos expériences internationales de l'utilisation de l'approche *Realist*

Les interventions de promotion de la santé et les inégalités sociales de santé

Au cours de nos expériences de recherches au Canada et en France concernant les interventions de promotion de la santé et leurs effets potentiels dans la réduction des inégalités sociales de santé, nous avons fait face à trois défis dans l'utilisation de l'approche *Realist*.

Le premier défi concerne la difficulté à disposer de l'élément contextuel dans les configurations CME. À la demande du ministère de la Santé du Canada, des chercheurs du Consortium canadien de recherche en promotion de la santé ont été mandatés pour évaluer l'efficacité des interventions en promotion de la santé (Hills, O'Neill, Carroll et MacDonald, 2004). Il s'agissait de construire un outil à caractère exploratoire afin d'évaluer l'efficacité des programmes et de mettre au jour les mécanismes potentiellement contributeurs de l'efficacité. Au stade de la réflexion sur le concept de mécanisme (voir *infra*) à cette époque de l'utilisation de l'approche *Realist*, les chercheurs évoquaient donc plutôt des stratégies d'intervention que des réactions cognitives des acteurs sociaux impliqués dans les programmes (tableau 2). Afin de concevoir un tel outil, nous avons réalisé plusieurs revues des écrits scientifiques afin de repérer certains mécanismes récurrents et potentiellement efficaces. Compte tenu de l'état de développement peu avancé de la recherche et des moyens disponibles à l'époque, les chercheurs se limitaient à mettre au jour les seuls mécanismes, laissant ainsi de côté les contextes et les effets, ce qui montre les limites d'une approche « budgétivore ». L'ambition des chercheurs était de découvrir ces mécanismes « candidats », donc potentiels, et, par la suite, de pouvoir en déceler dans d'autres contextes (Guichard et Ridde, 2010 ; Hills, Carroll et O'Neill, 2004).

Au cours de la recension des écrits francophones, nous avons très vite éprouvé des difficultés dans l'application de la démarche *Realist*. Nous ne reviendrons pas sur les problèmes habituels inhérents à l'accès à des revues francophones qui publient rarement de telles évaluations. L'utilisation de l'approche *Realist* a été ardue principalement du fait de la manière dont les us et coutumes scientifiques ont façonné les formes d'écriture. Outre la difficulté évoquée ci-dessus d'obtenir des informations sur les contextes, la nécessité de publier (« *publish or perish* ») a de son côté contribué au « démembrement » de nombreuses études. Ainsi, il est fréquent que les

évaluations des effets des interventions et les analyses de processus fassent l'objet de publications séparées, ce qui rend difficile leur arrimage. La seconde difficulté est liée au fait que les analyses des interventions dont il s'agit de faire une synthèse n'ont pas été rédigées dans ce but. Les cadres d'analyse et les théories à l'origine des travaux évaluatifs ne reprennent pas forcément, loin s'en faut, les éléments de l'approche *Realist*. Ainsi, la synthèse ne peut-elle découvrir ou inférer que peu d'éléments des écrits dont le contenu n'a pas été produit selon l'approche *Realist*. Voilà pourquoi des auteurs ayant récemment tenté une telle expérience en arrivent au même constat que nous: « Cette revue [...] doit être complétée par des évaluations utilisant l'approche réaliste afin de construire et de tester les théories de programme » (Dieleman *et al.*, 2009).

À la suite de cette recherche canadienne, des travaux effectués en France ont révélé un deuxième défi lié à l'application de l'approche *Realist*, celui de disposer de données sur les effets (E) des interventions. Nous avons fait valoir ailleurs que, dans le domaine de la promotion de la santé, une intervention est qualifiée d'efficace si elle est à la fois efficace et équitable, c'est-à-dire qu'elle contribue à réduire les inégalités sociales de santé (Ridde, 2007). Ainsi, nous avons choisi de mener une étude de cas multiples en France afin de collecter des données empiriques originales. Dans un premier temps, un guide d'entrevues qualitatives a été développé, reprenant les principaux mécanismes mis au jour au Canada, auquel ont été ajoutés et adaptés des mécanismes relevés dans deux autres outils employés dans des travaux hollandais (*Preffi*) et européen (*Closing the Gap*) (Guichard et Ridde, 2010). Dans un second temps, ce guide a servi de support à des entrevues en profondeur qui visaient à rendre compte des configurations CME. Il était évidemment impossible de passer en revue les 51 mécanismes de la grille. C'est pourquoi les configurations CME ont été examinées en vue d'isoler les mécanismes jugés par les acteurs propres à réduire les inégalités sociales de santé. Le défi majeur a été de trouver le E, car aucun des responsables des neuf interventions retenues n'avait organisé un système d'évaluation permettant de déterminer si l'effet visé, à savoir la réduction des inégalités sociales de santé, avait été atteint. Ainsi, nous avons pris comme mesure d'effet la *perception de la réduction* des inégalités par les intervenants. Cette perspective émique, pour reprendre un terme de l'anthropologie (Olivier de Sardan, 2008), constitue certes une limite scientifique au sens positiviste du terme.

Toutefois nous nous inscrivons dans une perspective constructiviste, où la construction du sens par les acteurs sociaux est centrale dans la production des connaissances. De plus, nous considérons cette étude comme une première étape dans la mise au jour d'un certain nombre de mécanismes causaux. À la fin de 2011, un programme de recherche ayant pour but de poursuivre ces travaux a démarré dans l'ouest de la France (www.iferiss.org/). Au-delà de cette limite liée à la perspective émiqque adoptée dans cette recherche, nous croyons que cette approche permet d'obtenir des données originales utiles pour apprécier l'efficacité et l'équité des actions.

Pendant, cette recherche a aussi placé les chercheurs devant un troisième défi, que laissaient entrevoir les difficultés déjà éprouvées : la définition du concept de mécanisme. En effet, depuis les travaux auxquels nous avons participé au Canada au début des années 2000, notre compréhension et notre utilisation du concept de mécanisme ont largement évolué, du fait des débats entre chercheurs (voir : www.itg.be/internet/ds/tde/index.html) et du besoin ressenti par certains de clarifier cette notion (Astbury et Leeuw, 2010). À la suite de cette application en France, nous avons compris qu'il fallait étudier plus en profondeur et de manière plus exhaustive tous les mécanismes présents dans différentes actions afin de relever certaines récurrences et de passer à des théories de moyenne portée. De plus, au moment où nous avons effectué cette recherche, nous n'avions pas eu le temps de développer une compréhension suffisamment précise de la notion de « mécanisme », appliquée aux interventions visant à lutter contre les inégalités sociales de santé. Cette notion est l'objet de débats depuis longtemps en sciences sociales (Hedström et Swedberg, 1996), mais ce n'est que récemment qu'elle a été explorée dans le domaine de l'évaluation des programmes (Astbury et Leeuw, 2010). Ainsi, plusieurs éléments qualifiés de mécanismes dans cette recherche (Ridde et Guichard, 2011) relevaient plus d'activités et de stratégies d'interventions (par exemple, la « mise en place de processus réflexifs et de dialogue critique ») que de raisonnements cognitifs ou d'attitudes et aptitudes des acteurs sociaux (par exemple, la « réflexion critique continue sur les processus et les résultats »). Nous devons donc poursuivre l'adaptation de la grille et l'identification des mécanismes en jeu pour être en mesure de comprendre leurs contributions à la réduction des inégalités sociales de santé et pouvoir l'appliquer à d'autres contextes. Ensuite, il conviendrait

évidemment d'obtenir des données sur le E de la famille $C + M = E$, c'est-à-dire les effets relatifs à la réduction des inégalités sociales de santé. Comme de telles évaluations sont rares, il faudrait tenter d'appliquer cette démarche dans des projets de recherche où les interventions visent spécifiquement à réduire les inégalités sociales de santé.

La modélisation des configurations organisationnelles en Afrique associée à l'évaluation Realist

Nous avons mené une évaluation *Realist* pour étudier les effets des programmes d'amélioration de la qualité des soins mis en œuvre dans les services publics de santé en Afrique durant les vingt dernières années (Blaise, 2005). Partant du postulat que « la qualité est le principe d'action unique autour duquel l'ensemble de l'organisation et de ses acteurs doivent s'organiser » (Minvielle *et al.*, 1994), nous avons étudié dans quelle mesure les projets avaient permis aux systèmes de santé d'adopter les principes du management de la qualité et d'inscrire de façon pérenne la qualité au cœur de leur management (théorie d'intervention). Nous nous sommes concentrés sur le processus de mise en œuvre du management de la qualité (mécanisme M) dont nous avons observé les effets sur l'organisation, la configuration, le fonctionnement, les acteurs et la « culture ». Ces éléments d'évaluation une fois précisés, nous avons mené une série d'études de cas qui nous a permis d'examiner des « configurations CME » dans des contextes multiples et variés (contexte C).

Au terme de cette démarche *Realist*, nous avons formulé une théorie de moyenne portée qui peut être résumée ainsi : « L'introduction de projets qualité valorisant la prise d'initiative et le changement n'ont pas transformé le management de systèmes de santé structurés comme des organisations mécanistes, hiérarchisées et coordonnées par la standardisation des procédures. Tolérées pour un temps dans un contexte de projet, les prises d'initiatives propres à la démarche qualité y sont inhibées. L'évolution attendue vers l'organisation apprenante ne s'est pas concrétisée. »

Tirant les leçons de cette expérience, nous recensons les obstacles que nous avons dû franchir pour mener à bien cette évaluation. Nous avons d'abord dû bien circonscrire le mécanisme étudié (M), puis le décrire dans chacune des études de cas, de façon à être certains qu'il s'agissait bien du même mécanisme à l'œuvre produisant ses effets de manière variable d'un

contexte à l'autre. Le mécanisme observé était la transformation du management des services de santé après la mise en œuvre des projets qualité. Ensuite, nous avons dû décrire et caractériser le contexte pour le comparer d'une étude de cas à l'autre. La modélisation des configurations organisationnelles proposée par Mintzberg (1982) nous a permis de bien cerner la structure et les modes de coordination des organisations.

Pour amorcer la démarche, il nous fallait formuler une théorie initiale. Cette théorie a été établie à partir d'une large comparaison internationale s'appuyant sur la modélisation de Mintzberg et a permis d'établir une typologie des organisations de soins, puis de comparer la mise en place et l'évolution du management de la qualité dans ces divers types d'organisation (Blaise et Kegels, 2004). Nous pensons qu'il faut faire montre de prudence lorsqu'on veut généraliser des interventions en s'appuyant sur les théories ainsi développées. Les lunettes indispensables pour réduire la complexité n'offrent qu'un angle de vue. Changer de « lunettes », d'outil de modélisation, offre une nouvelle perspective sur les mêmes phénomènes. En modifiant la perspective, la perception change; elle n'efface pas la précédente, mais la complète et l'enrichit. Si on poursuit la métaphore, en tournant autour de l'objet d'étude, il apparaît lui-même en transformation permanente, de sorte qu'il pourrait se présenter différemment lorsqu'on revient au point d'observation initial au cours du processus d'analyse. Il s'agit donc d'un processus continu. Il ne s'agit pas de comprendre le fonctionnement d'un mécanisme stable en examinant successivement chacune des pièces, comme on désassemblerait une horloge; il s'agit de comprendre des dynamiques en constante évolution et ainsi de donner du sens à l'action des acteurs afin d'inspirer et d'orienter des interventions futures. Le caractère partiel et provisoire des théories témoigne non pas du caractère inadéquat de la méthode, mais plutôt de la nature complexe du système dans lequel ont lieu les interactions. En conséquence, si l'évaluation *Realist* assure une validité externe en rapprochant les études de cas, les théories de moyenne portée qu'elle permet de développer restent incomplètes et provisoires, et le processus d'évaluation est par conséquent cyclique et continu.

**Évaluer des interventions a posteriori :
du caractère applicable de l'évaluation *Realist***

L'approche *Realist* a été utilisée pour analyser deux interventions dont l'objectif était d'améliorer les compétences des médecins communautaires au Mali (Coulibaly *et al.*, 2007; Van Dormael *et al.*, 2008; Van Dormael, Dugas et Diarra, 2007), l'une portant sur la formation professionnelle, et l'autre sur l'échange Nord-Sud. Chaque intervention a été considérée comme une étude de cas exploratoire en vue de développer une théorie. La théorie initiale était basée sur les hypothèses (implicites) formulées par les participants. Une analyse des documents de projet, d'articles publiés et d'une entrevue avec le chercheur principal a conduit à mettre au jour des configurations CME, lesquelles ont amené à raffiner la théorie.

Sur le plan conceptuel, le principal obstacle de l'approche est l'intégration du contexte à la théorie. La configuration CME est une description de l'étude de cas individuelle, une explication du passé. L'objectif ultime de l'approche *Realist* est donc de formuler une théorie de moyenne portée en des termes plus généraux. Cependant, saisir la multiplicité des différents contextes et mécanismes dans une théorie de moyenne portée est un véritable jeu d'équilibriste. Ainsi, comment atteindre un niveau suffisant de précision pour qu'une déclaration soit crédible (« *trustworthy* ») et en même temps ait un niveau suffisant de généralisation théorique pour qu'elle soit utile dans d'autres situations? Nous pensons que l'utilité de l'évaluation *Realist* réside essentiellement dans l'explication du passé basée sur les configurations CME.

Sur le plan pratique, le premier défi a été de délimiter les éléments à analyser et de distinguer ce qui relevait du contexte, des mécanismes et des effets. La chaîne d'événements (non linéaire) de la plupart des interventions est composée de différentes strates et étapes, et une classification simple de type CME ne permet pas de décrire le processus à l'œuvre de manière satisfaisante. En effet, des exemples d'éléments additionnels de cette chaîne sont l'intervention elle-même et les effets intermédiaires, qui pourraient être à leur tour considérés comme des mécanismes conduisant aux effets ultimes. Dans cette expérience, nous avons étudié comment les différentes caractéristiques de l'intervention – la formation professionnelle – pouvaient améliorer l'attitude des médecins et leurs compétences.

Un autre défi a consisté à relier les mécanismes spécifiques et les éléments du contexte à un effet. Dans un système complexe, il est impossible

de démêler les différentes composantes d'une intervention pour rendre compte de leur effet individuel. Par exemple, l'amélioration de l'attitude des médecins est-elle causée par l'échange avec des modèles, l'amélioration de l'estime de soi ou d'autres facteurs ?

Enfin, l'évaluation est souvent réalisée après l'intervention, ce qui pose le problème de la formulation *a posteriori* de la théorie. En effet, si la théorie est formulée après que l'intervention a eu lieu, la vision du chercheur est biaisée par ce qu'il observe dans les faits. Il risque ainsi de se concentrer sur l'analyse des déterminants qui semblent évidents *a posteriori* et d'en négliger d'autres. Les limites pratiques à de nombreuses évaluations empêchent de réaliser plusieurs cycles d'analyse qui permettraient de développer la théorie et de la vérifier. Pourtant, idéalement, l'évaluation *Realist* est un processus longitudinal qui débute au moment de la planification de l'intervention.

Conclusion

Adopter une démarche *Realist* pour mieux appréhender la complexité des interventions sociales est une approche intellectuellement séduisante. Nos expériences, comme celles d'autres auteurs qui ont appliqué ces méthodes, montrent que l'enthousiasme suscité par son utilisation ne peut faire abstraction des défis méthodologiques qu'elles comportent. Le premier défi concerne les revues *Realist* et l'absence dans les articles scientifiques d'informations susceptibles d'aider à formuler les théories des interventions et à développer les configurations CME, dans un contexte où l'espace de publication est restreint et où les évaluations qui utilisent l'approche *Realist* sont encore peu nombreuses. Le second défi consiste à faire la distinction entre les éléments relevant du contexte et ceux relevant du mécanisme (dont la conceptualisation est complexe et mériterait un examen plus poussé), voire des effets, pour la construction des configurations CME. Il s'agit en effet de déterminer clairement le niveau d'analyse permettant d'observer les régularités des configurations CME qui intéressent le chercheur. Il s'agit d'un exercice difficile car il faut s'ajuster pour à la fois obtenir un niveau de précision suffisant pour appréhender l'intervention et tester la théorie de moyenne portée. Enfin, le défi qui consiste à disposer du temps nécessaire pour maîtriser l'approche, atteindre la maturité suffisante pour pouvoir analyser les données et déterminer ce que sont les mécanismes dans le

contexte de l'étude semble aujourd'hui remettre en question son utilité pour informer la prise de décision sur des problématiques urgentes. Toutefois, si l'approche *Realist* est, comme toute innovation méthodologique, difficile d'emploi, elle n'en reste pas moins un outil conceptuel prometteur dans le cadre de la recherche évaluative.

Références

- ALKIN, M. C. (dir.) (2004), *Evaluation Roots. Tracing Theorists' Views and Influences*, Thousand Oaks, Sage.
- ASTBURY, B. et F. L. LEEUW (2010), Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation, *American Journal of Evaluation*, 31(3), 363-381.
- BLAISE, P. (2005), *Culture, qualité et organisation bureaucratique. Le défi du changement dans les systèmes publics de santé. Une évaluation réaliste de projets qualité en Afrique*, Thèse de doctorat, Université Libre de Bruxelles, Bruxelles.
- BLAISE, P. et G. KEGELS (2004), A realistic approach to the evaluation of the quality management movement in health care systems: a comparison between European and African contexts based on Mintzberg's organizational models. *International Journal of Health Planning and Management*, 19(4), 337-364. doi:10.1002/hpm.769.
- BLAISE, P., B. MARCHAL, P. LEFÈVRE et G. KEGELS. (2010), Au-delà des méthodes expérimentales, l'approche réaliste en évaluation. Dans L. POTVIN, M. J. MOQUET et C. JONES (dir.), *Réduire les inégalités sociales de santé* (p. 285-296), Saint-Denis, INPES, coll. « Santé en action ».
- BYNG, R., I. NORMAN et S. REDFERN (2005), Using Realistic Evaluation to Evaluate a Practice-level Intervention to Improve Primary Healthcare for Patients with Long-term Mental Illness. *Evaluation: The International Journal of Theory, Research and Practice*, 11(1), 69-93. doi:10.1177/1356389005053198.
- BYNG, R., I. NORMAN, S. REDFERN et R. JONES (2008), Exposing the key functions of a complex intervention for shared care in mental health: case study of a process evaluation, *BMC Health Services Research*, 8, 274. doi:10.1186/1472-6963-8-274.
- CHEN, H. T. (2005), *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness*, Thousand Oaks, Sage Publications.
- CLAMP, S. et J. KEEN (2007), Electronic health records: Is the evidence base any use? *Informatics for Health and Social Care*, 32(1), 5-10. doi:10.1080/14639230601097903.

- CLARK, A., H. K. WHELAN, R. BARBOUR et P. D. MACINTYRE (2005), A realist study of the mechanisms of cardiac rehabilitation, *Journal of Advanced Nursing*, 52(4), 362-371.doi:10.1111/j.1365-2648.2005.03601.x.
- DIELEMAN, M., B. GERRETSEN et G. J. VAN DER WILT (2009), Human resource management interventions to improve health workers' performance in low and middle income countries: a realist review, *Health Research Policy and System*, 7(7).doi:10.1186/1478-4505-7-7.
- DONALDSON, S. I. (2007), *Program Theory-driven Evaluation Science: Strategies and Applications*, New York, Lawrence Erlbaum Associates.
- EVANS, D., et A. KILLORAN (2000), Tackling health inequalities through partnership working: learning from a realistic evaluation, *Critical Public Health*, 10(2), 125-140.doi:10.1080/09581590050075899.
- GRANT, M. J. et A. BOOTH (2009), A typology of reviews: An analysis of 14 review types and associated methodologies, *Health Information et Libraries Journal*, 26(2), 91-108.
- GREENHALGH, T., E. KRISTJANSSON et V. ROBINSON (2007), Realist review to understand the efficacy of school feeding programmes, *British Medical Journal*, 335, 858-861.doi:10.1136/bmj.39359.525174.AD.
- GUICHARD, A. et V. RIDDE (2010), Une grille d'analyse des actions pour lutter contre les inégalités sociales de santé. Dans L. POTVIN, M. J. MOQUET et C. JONES (dir.), *Réduire les inégalités sociales de santé* (p. 297-312), Saint-Denis, INPES, coll. « Santé en action ».
- HEDSTRÖM, P. et R. SWEDBERG (1996), Social Mechanisms, *Acta Sociologica*, 39, 281-308.
- HILLS, M., M. O'NEILL, S. CARROLL et M. MACDONALD (2004), Efficacité des interventions communautaires visant à promouvoir la santé: un instrument d'évaluation. Rapport de recherche présenté à Santé Canada par le Consortium canadien pour la recherche en promotion de la santé.
- HILLS, M. D., S. CARROLL et M. O'NEILL (2004), Vers un modèle d'évaluation de l'efficacité des interventions communautaires en promotion de la santé: compte-rendu de quelques développements nord-américains récents, *Promotion et Éducation - Hors série 1*, 11, 17-21.doi:10.1177/10253823040110010105.
- JACKSON, L., L. LANGILLE, L. LYONS, J. HUGHES, D. MARTIN et V. WINSTANLEY (2009), Does moving from a high-poverty to lower-poverty neighborhood improve mental health? A realist review of «Moving to Opportunity», *Health et Place*, 15(4), 961-970.doi:10.1016/j.healthplace.2009.03.003.
- JULNES, G., M. M. MARK et G. T. HENRY (1998), Promoting realism in evaluation: Realistic evaluation and the broader context, *Evaluation*, 4(4), 483-504. doi:10.1177/13563899822208743.
- KANE, S. S., B. GERRETSEN, R. SCHERPBIER, M. DAL POZ et M. DIELEMAN (2010), A realist synthesis of randomised control trials involving use of community

- health workers for delivering child health interventions in low and middle income countries, *BMC Health Services Research*, 10, 286. doi: 10.1186/1472-6963-10-286. Repéré à www.biomedcentral.com/1472-6963/10/286
- KENNEDY, A., A. ROGERS et C. GATELY (2005), Assessing the introduction of the expert patients programme into the NHS: A realistic evaluation of recruitment to a national lay-led self-care initiative, *Primary Health Care Research and Development*, 6(2), 137-148. doi: 10.1191/1463423605pc2410a.
- LAGARDE, M. et N. PALMER (2006), *Evidence from systematic reviews to inform decision making regarding financing mechanisms that improve access to health services for poor people: A policy brief prepared for the International Dialogue on Evidence-Informed Action to Achieve Health Goals in Developing Countries (IDEAHealth)*, Genève, Alliance for Health Policy and Systems Research. Repéré à www.who.int/rpc/meetings/HealthFinancingBrief.pdf
- LAW, J., C. PLUNKETT, J. TAYLOR et M. GUNNING (2009), Developing policy in the provision of parenting programmes: integrating a review of reviews with the perspectives of both parents and professionals, *Child: Care, Health and Development*, 35(3), 302-312. doi: 10.1111/j.1365-2214.2009.00939.x.
- LEEMAN, J., Y. K. CHANG, E. J. LEE, C. I. VOILS, J. CRANDELL et M. SANDELOWSKI, (2010), Implementation of antiretroviral therapy adherence interventions: A realist synthesis of evidence, *Journal of Advanced Nursing*, 66(9), 1915-1930. doi: 10.1111/j.1365-2648.2010.05360.x.
- MARCHAL, B., M. DEDZO et G. KEGELS (2010), A realist evaluation of the management of a well-performing regional hospital in Ghana, *BMC Health Services Research*, 10(24). doi: 10.1186/1472-6963-10-24. Repéré à www.biomedcentral.com/1472-6963/10/24
- MCCORMACK, B., J. WRIGHT, B. DEWAR, G. HARVEY et K. BALLANTINE (2007), A realist synthesis of evidence relating to practice development: methodology and methods, *Practice Development in Health Care*, 6(1), 5-24. doi: 10.1002/pdh.210.
- MERTON, R. K. (1968), *Social Theory and Social Structure*, New York, Free Press.
- MINTZBERG, H. (1982), *Structure et dynamique des organisations*, Paris, Éditions d'organisation.
- MINVIELLE, E., M. BREMOND, A. DEVELAY, P. LOMBRIL et M. BRODIN (1994), Vivent les « démarches qualité » à l'hôpital, *Gestions hospitalières*, 340, 727-738.
- MORIN, E. (2005), *Introduction à la pensée complexe*, Paris, Seuil.
- O'CAMPO, P., M. KIRST, N. SCHAEFER-MCDANIEL, M. FIRESTONE, A. SCOTT et K. MCSHANE (2009), Community-based services for homeless adults experiencing concurrent mental health and substance use disorders: A realist approach to synthesizing evidence, *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 86(6), 965-989. doi: 10.1007/s11524-009-9392-1.

- OLIVIER DE SARDAN, J.-P. (2008), *La rigueur du qualitatif. Les contraintes empiriques de l'interprétation socio-anthropologique*, Louvain-la-Neuve, Academia Bruylant.
- OROVIOGOICOECHEA, C. et R. WATSON (2009), A quantitative analysis of the impact of a computerised information system on nurses' clinical practice using a realistic evaluation framework, *International Journal of Medical Informatics*, 78(12), 839-849. doi:10.1016/j.ijmedinf.2009.08.008.
- PAWSON, R. (2002), Evidence-based policy: The promise of « Realist Synthesis ». *Evaluation*, 8(3), 340-358. doi:10.1177/135638902401462448.
- PAWSON, R. (2003), Nothing as practical as good theory, *Evaluation*, 9(4), 471-490. doi:10.1177/1356389003094007.
- PAWSON, R., T. GREENHALGH, G. HARVEY et K. WALSH (2004), Realist synthesis: An introduction (RMP Methods Paper 2/2004), *ESRC Research Methods Programme*, 2. Repéré à www.ccsr.ac.uk/methods/publications/documents/RMPmethods2.pdf
- PAWSON, R. et N. TILLEY (1997), *Realistic Evaluation*, Londres, Thousand Oaks, Sage.
- PLUYE, P., L. NADEAU, M.-P. GAGNON, R. GRAD, J. JOHNSON-LAFLEUR et F. GRIFFITHS (2009), Les méthodes mixtes pour l'évaluation des programmes. Dans V. RIDDE et C. DAGENAIS (dir.), *Approches et pratiques en évaluation de programmes* (p. 123-141), Montréal, Presses de l'Université de Montréal.
- POPE, C., N. MAYS et J. POPAY (2007), *Synthesizing Qualitative and Quantitative Health Evidence: A guide to Methods*, Maidenhead, McGraw-Hill Education.
- RIDDE, V. (2007), Reducing social inequalities in health: public health, community health or health promotion?, *Promotion & Education*, 14(2), 63-67.
- RIDDE, V. et C. DAGENAIS (2009), *Théories et pratiques en évaluation de programmes: Manuel d'enseignement*, Montréal, Presses de l'Université de Montréal.
- RIDDE, V., T. DELORMIER et L. GAUDREAU (2007), Evaluation of empowerment and effectiveness: Universal concepts? Dans D. McQUEEN et C. JONES (dir.), *Global Perspectives on Health Promotion Effectiveness*, New York, Springer / IUHPE.
- RIDDE, V. et A. GUICHARD (2011), Perception de quelques mécanismes favorables à la réduction des inégalités sociales de santé en France, *Global Health Promotion*, 18(3), 47-60.
- ROBERT, E., V. RIDDE, B. MARCHAL et P. FOURNIER (2012), Protocol: a realist review of user fee exemption policies for health services in Africa, *BMJ Open*, 2(1), e000706.
- ROBERT, E., V. RIDDE, P. FOURNIER et A. BICABA (2009), How to use the Realist approach to inform policy-making: The example of a research method on user fees abolition policies in Africa. Evidence in Uncertain Times: Meeting

the challenge; 19-20 mai 2010; Ottawa (Canada): 8th Annual Symposium of Canadian Cochrane Centre.

- RYCROFT-MALONE, J., M. FONTENLA, D. BICK et K. SEERS (2010), A realistic evaluation: The case of protocol-based care, *Implementation Science*, 5(38).
- WILSON, V., B. MCCORMACK et G. IVES (2008), Developing healthcare practice through action learning: Individual and group journeys, *Action Learning: Research and Practice*, 5(1), 21-38. doi:10.1080/14767330701880226.
- WONG, G., T. GREENHALGH, et R. PAWSON (2010), Internet-based medical education: A realist review of what works, for whom and in what circumstances, *BMC Medical Education*, 10(1). doi: 10.1186/1472-6920-10-12.
- YIN, R. K. (1994), *Case Study Research: Design and Methods* (2^e éd.), Londres, New Delhi, Sage Publications.

DEUXIÈME PARTIE

PRATIQUES ÉVALUATIVES

15

Une évaluation participative avec renforcement du pouvoir d'agir

Valéry Ridde

Depuis quelques années, les organismes communautaires des régions rurales du Québec tentent de mettre en œuvre des programmes de « travail de rue » destinés aux jeunes. Devant les requêtes des organismes subventionnaires et face à la concurrence créée par l'insuffisance des ressources disponibles, ces projets doivent faire la démonstration de leur efficacité. Or, il faut bien reconnaître que les évaluations de tels projets ne sont pas légion. L'une des hypothèses que nous aimerions émettre à cet égard découle du fait qu'il est délicat de faire appel aux stratégies traditionnelles d'évaluation de programmes pour ce type de projets. Leur nature particulière, les valeurs défendues par les travailleurs de rue (TR) ainsi que l'intervention auprès de jeunes directement dans leur milieu de vie ne rendent pas le processus d'évaluation facile. Dans ce chapitre*, nous tentons de montrer que le recours à une démarche participative de type

* Ce chapitre est adapté d'un article paru dans *Service social* (Ridde, Baillargeon *et al.*, 2003). Nous remercions l'ensemble de l'équipe de Ressource alternative des jeunes de Bellechasse, les travailleurs de rue ainsi que les jeunes participants, sans qui ce texte n'aurait pu voir le jour.

empowerment est l'une des solutions envisageables pour l'évaluation des projets de travail de rue.

Le contexte

Le contexte socioéconomique

Le projet dont il est ici question a été élaboré et piloté par l'organisme Ressource alternative des jeunes de Bellechasse (RAJB). La mission de cet organisme est de favoriser l'intégration sociale, scolaire, professionnelle et économique des jeunes de la Municipalité régionale de comté (MRC) de Bellechasse. Le projet « Travail de rue » a été conçu en continuité avec une intervention de même type démarrée au milieu de l'année 1996. Le financement renouvelé du projet a permis d'étendre les activités à cinq municipalités supplémentaires et d'envisager l'évaluation de l'ensemble des interventions.

La MRC de Bellechasse est située dans la région Chaudière-Appalaches. Les habitants de cette région sont moins scolarisés que la moyenne provinciale, ils sont moins nombreux à être au chômage, mais disposent de revenus moins élevés et ont un taux ajusté de mortalité par suicide plus important. Plaçant la région au deuxième rang à l'échelle provinciale, le taux de mortalité par suicide est en hausse croissante depuis 15 ans, notamment pour les hommes, et surtout pour les jeunes (INSPQ, 2001). Cette région regroupe 19 municipalités de 406 à 3160 personnes. Les jeunes de 15 à 24 ans représentent 13,2 % de sa population, soit 3910 personnes. Entre 1991 et 1996, on a pu observer un phénomène d'exode des jeunes de 15 à 19 ans bien plus important à Bellechasse que dans le reste de la région. Le pourcentage de jeunes diplômés après sept ans de scolarité au secondaire est inférieur à ce qui a été constaté dans l'ensemble de la région (Fafard *et al.*, 2000).

Le travail de rue au Québec

Au Québec, le travail de rue est apparu à la fin des années 1960. Après un passage à vide, dans les années 1970, les travailleurs de rue étant pour la plupart intégrés dans le réseau des affaires sociales, il a repris de l'ampleur depuis les années 1980. À l'origine, il s'adressait principalement aux per-

sonnes toxicomanes de Montréal. Puis, les problématiques prises en charge se sont multipliées, et l'émergence du phénomène des jeunes de la rue a favorisé la création de plusieurs organismes, au début des années 1990, ce qui a conduit à la création, en 1993, de l'Association des travailleurs et travailleuses de rue du Québec (ATTRueQ). C'est seulement à partir de cette date que quelques expériences ont été menées en région rurale (Huterbise *et al.*, 2000; Beaulé et Simard, 2002). Deux facteurs essentiels seraient responsables du renouveau des programmes pour les jeunes de la rue (Duval et Fontaine, 2000; Parazelli, 2002). D'abord, l'accroissement du nombre de jeunes dans la rue à la suite de changements économiques et de l'augmentation du décrochage scolaire, puis la volonté des acteurs institutionnels, à la suite de l'instauration par le gouvernement du fameux « virage ambulatoire », de se rapprocher des individus et d'agir plus souvent dans les milieux de vie. Évidemment, il faut ajouter à cela que les actions entreprises dans le cadre du travail de rue sont en lien très étroit avec les différentes problématiques inscrites dans ces politiques publiques (violence, maladies transmissibles, alcool, santé mentale, etc.).

Il faut d'emblée souligner que le travail de rue n'est pas compris de la même manière par tous. Il en existe de multiples définitions et la plus grande confusion règne encore aujourd'hui à cet égard. Une myriade d'expressions sont employées pour y faire référence : *outreach* ou travail de proximité, de rue, de corridor, de milieu, de quartier, etc. Duval et Fontaine (2000) ont conçu une typologie des actions du travail de rue caractérisée par deux approches : normative et autonome. La première correspond à des activités conçues en guise de prolongement aux interventions institutionnelles, le TR jouant un rôle de médiateur. Cette approche est encore dominante, notamment pour les organismes publics, puisqu'elle n'est pas incompatible avec des interventions de promotion de la santé. Elle conçoit les jeunes comme des « groupes à risque », des sujets identifiés par une « santé publique épidémiologique », des individus dont il faut changer les comportements malsains. Selon Parazelli (2002, p. 300), c'est la « *pression gestionnaire* » qui serait responsable de cette dérive interventionniste de type normatif. La seconde approche vise à rendre les jeunes plus autonomes et plus responsables par l'intermédiaire d'une action éducative globale fondée sur leur émancipation. Au Québec, l'ATTRueQ promeut cette deuxième approche, qui est à la fois centrée sur les besoins des jeunes et soucieuse de leur environnement et de leur mode

de vie. Il faut par contre souligner que ce type d'interventions est relativement rare.

Les stratégies d'évaluation du travail de rue

Les évaluations des projets de travail de rue au Québec sont rares. La plupart d'entre elles ont porté sur les processus d'intervention et peu, sinon aucune, sur leurs effets (Simard *et al.*, 2000). Celles qui visaient à porter un jugement sur leur efficacité se sont surtout penchées sur la problématique des MST/sida (Moore *et al.*, 1998).

Parazelli (2002) soutient qu'il est délicat de vouloir définir clairement le concept de « jeune de la rue » et que, pour éviter les interprétations abusives, il convient de laisser une place importante à l'analyse de ce phénomène par les jeunes eux-mêmes. En effet, tel que le rapportent Fortier et Roy dans leur revue de littérature citée par Parazelli (p. 298) : « Peu d'articles [...] font état de ce que pensent les jeunes eux-mêmes. » À notre connaissance, une seule recherche évaluative en milieu rural québécois a été organisée en prenant en compte le point de vue des jeunes (Simard *et al.*, 2002). Cependant, bien que cette étude ait été menée en collaboration avec les principaux intéressés, les chercheurs ont procédé eux-mêmes à la collecte et à l'analyse des données. La même approche a été expérimentée dans un projet de recherche-action participative avec les jeunes de la rue à Toronto (Poland *et al.*, 2002). Enfin, une expérience fort intéressante s'est déroulée à Ottawa, où les intervenants et les jeunes de la rue ont participé à l'évaluation (Whitmore et McKee, 2001).

La description du projet

À la suite de l'intégration des travailleurs de rue dans la MRC de Bellechasse, en 1998, l'organisme RAJB a ressenti le besoin d'assurer une présence préventive continue dans les milieux de vie des jeunes de 12 à 25 ans. En ce qui concerne la finalité de ce projet, à la lumière des entretiens réalisés avec les TR et en regard de leur philosophie de pratique, le travail de rue est défini selon une approche visant à favoriser la prise de pouvoir par les jeunes, l'autonomie et l'accompagnement étant des concepts centraux de cette approche. Laissant de côté les stratégies instrumentales, le travail de rue tel qu'il est pratiqué dans le cadre de ce projet est en phase avec les valeurs

véhiculées par l'ATTRueQ et s'apparente tant au mode d'action de l'accompagnement (Parazelli, 2002) qu'à l'approche favorisant l'autonomie des jeunes (Duval et Fontaine, 2000), qui sont considérés comme des acteurs sociaux, comme des êtres politiques. L'intervention des TR vise donc l'habilitation et l'*empowerment* des jeunes. Elle est globale, les jeunes ne sont pas « découpés » par problématique (suicide, toxicomanie, etc.) et les TR cherchent à répondre aux besoins exprimés dans leur globalité. Le projet de Bellechasse ne s'intéresse pas uniquement aux jeunes marginaux, mais à l'ensemble des jeunes du territoire.

En ce qui a trait aux différents objectifs à long terme que s'étaient fixés les responsables de ce projet, il s'agissait de contribuer à la promotion de la santé et du bien-être des jeunes. À moyen terme, les effets attendus consistaient à favoriser le renforcement du pouvoir des jeunes pour qu'ils soient en mesure d'agir sur les déterminants de leur santé. Au cœur de ce projet figurent donc les concepts d'*empowerment*, d'habilitation et d'autonomie. Pour y arriver, les TR ont conçu quatre objectifs à court terme qui devaient leur permettre d'observer des résultats tout au long du processus de mise en œuvre du projet. Signalons que ces objectifs ont été définis plus précisément à la suite de réunions de travail avec les TR, lors de la phase préévaluative (voir chapitre 1). Ces objectifs sont les suivants :

- *Dépister.* Les TR, par leur intégration progressive dans le milieu, devront apprendre à connaître les jeunes de leur secteur et être en mesure de reconnaître ceux pour lesquels il sera nécessaire de mettre une intervention spécifique en œuvre.
- *Informers.* Lorsque les jeunes désirant ou nécessitant un soutien seront identifiés, les TR s'attacheront à répondre à leurs requêtes et à leur donner les informations dont ils auront besoin.
- *Orienter.* Plus tard, dans la démarche d'intervention, certains jeunes devront chercher du soutien auprès d'autres intervenants que les TR. Ces derniers joueront dans ce cas le rôle d'intermédiaires, de *between*, disent-ils, et interviendront pour orienter les jeunes vers les ressources les plus en mesure de répondre à leurs besoins du moment.
- *Responsabiliser.* Cet objectif vise à faire en sorte que les jeunes soient capables de prendre des décisions qui les concernent en toute connaissance de cause. Il s'agit du résultat à court terme le plus poussé, au regard du continuum des interventions des TR.

Pour rendre plus explicites les objectifs visés, afin de ne pas nous tromper lors de la phase évaluative, nous avons demandé aux TR de nous donner leurs propres définitions des termes qui les résument (tableau 15.1).

TABLEAU 15.1**Définitions des objectifs selon les travailleurs de rue**

Objectifs	Définitions des travailleurs de rue
Dépister	Toutes les observations qui permettent d'établir une liste des rites, du rythme, des cultures, des croyances et des différents réseaux sociaux.
Informar	Toutes les discussions qui fournissent de l'information sans que cela soit nécessairement thérapeutique.
Orienter	Toute intervention qui touche à la référence.
Responsabiliser	Tous les gestes qui englobent l'accompagnement, la distribution de matériel et l'intervention en situation de crise.

Sur le plan de l'intervention, les TR se sont répartis sur le territoire en fonction de trois grandes régions. Trois équipes devaient intervenir sur le terrain : une composée d'hommes seulement, et une autre, de femmes, alors qu'une dernière était mixte. En ce qui a trait aux activités, il faut préciser que l'atteinte des objectifs exposés plus haut n'était possible, selon les promoteurs du projet eux-mêmes, que dans la mesure où l'intégration des TR dans le milieu de vie des jeunes se concrétisait. Quatre étapes ont donc jalonné les actions mises en œuvre par les intervenants et ont raisonnablement pu être organisées dans l'échelle temporelle indiquée : l'observation (6 à 9 mois), l'intégration (9 à 12 mois), l'implication (12 à 16 mois) et l'intervention (16 à 24 mois). Ces étapes ne constituent pas un processus linéaire, elles peuvent être simultanées et concourir à l'atteinte des quatre objectifs énoncés.

Le processus d'évaluation

Comme nous l'avons mentionné plus haut, rares sont les évaluations de projets de travail de rue, au Québec, qui ont porté sur leurs effets (Simard *et al.*, 2000). Il faut donc reconnaître, d'une part, l'ampleur du défi que représentait le mandat d'évaluer les effets du projet à Bellechasse et, d'autre part, le fait qu'il est indispensable, compte tenu du contexte et de la nature de ce type d'interventions, d'obtenir le point de vue des jeunes et d'impliquer les intervenants dans une évaluation participative émancipatrice.

L'approche évaluative

Une stratégie évaluative participative reposant sur une approche pluraliste fondée sur la négociation est qualifiée, avec d'autres (« *responsive, naturalistic, illuminative, utilization-focused, adversial* »), d'évaluation de quatrième génération, si l'on se situe dans un continuum historique (Guba et Lincoln, 1987, p. 208). Elle permet aux acteurs de participer activement à toutes les étapes de l'évaluation, contribuant ainsi à établir un climat de confiance tout en leur permettant d'acquérir de nouvelles compétences propres à l'évaluation de programmes. Le rôle des évaluateurs externes est d'offrir aux différentes parties intéressées une expertise et un accompagnement tout au long de ce processus. Dans ce domaine de l'évaluation participative, Cousins et Whitmore (1998) soutiennent que deux courants de pensée existent. D'un côté, il y a l'approche dite pratique, ou orientée sur l'utilisation, qui permet aux organisations de prendre des décisions ou de résoudre des problèmes particuliers. De l'autre, on retrouve l'approche émancipatrice, centrée sur l'*empowerment*, où la participation des acteurs vise à la fois l'autonomie des intervenants dans l'évaluation de leur travail et un changement social favorisé par le transfert de connaissances et de pouvoirs aux principaux intéressés¹.

Assurément, la stratégie évaluative participative la plus appropriée pour le projet de Bellechasse est l'approche émancipatrice, et ce, pour deux raisons principales. D'une part, l'accompagnement dans l'évaluation est essentiellement organisé à l'intention des TR (pour leur plus grande autonomie dans l'évaluation de leur travail) et non des jeunes ou de la communauté. D'autre part, le changement social souhaité vise la responsabilisation des jeunes. Préciser cela permet de répondre aux critiques émises contre les tenants de l'évaluation émancipatrice voulant que ne soit jamais clairement précisé qui est vraiment émancipé (« *who is really empowered* ») (Scriven, 1997, p. 168).

Plusieurs raisons peuvent être évoquées pour justifier l'emploi de l'approche émancipatrice (Ridde *et al.*, 2003). Dans ce chapitre, on se limitera à dire qu'il est possible de faire un parallèle étroit entre, d'une part, les deux

1. Dans un autre article (Ridde, 2006b), nous analysons (et critiquons la pertinence de) la proposition de ces mêmes auteurs d'ajouter à ces approches un troisième courant, soit un courant pluraliste ou *stakeholder-based evaluation* (Weaver et Cousins, 2005).

stratégies d'intervention du travail de rue (normative et autonome) et, d'autre part, les deux courants de pensée concernant l'évaluation participative (pratique et émancipatrice). Justement parce que le projet de Belle-chasse repose sur des interventions visant l'émancipation et l'*empowerment* des jeunes (stratégie de l'autonomie), il paraît d'autant plus pertinent d'avoir recours à une évaluation participative de type *empowerment*.

Le type d'évaluation et les outils employés

L'objectif global de l'évaluation de ce projet est de mesurer le degré d'atteinte des effets visés à court terme, c'est-à-dire de porter un jugement sur son efficacité. Cinq outils complémentaires ont été utilisés, mais nous ne nous attarderons ici que sur leurs éléments qui permettent de mieux comprendre la démarche évaluative émancipatrice s'étalant sur une période de 14 mois.

- Plusieurs réunions ont été organisées dans les locaux des promoteurs du projet. Elles ont été utiles pour préciser les objectifs du projet, pour planifier les activités de l'évaluation et pour effectuer les discussions nécessaires à la préparation des instruments de collecte de données. Une réunion animée selon les principes d'un groupe de discussion a été organisée à la fin de l'évaluation pour recueillir de nouvelles informations et revenir sur la démarche évaluative.
- Les documents étudiés ont été les propositions de projets, les rapports d'activités, les mémos écrits par les travailleurs de rue, etc.
- Avant le démarrage de l'évaluation, les responsables du projet avaient créé une grille de statistiques mensuelles. Au regard des objectifs du projet à évaluer, et considérant que celui-ci était financé par le ministère de la Santé du Canada (Santé Canada) et qu'il faisait directement référence à une intervention sur les déterminants de la santé, il a été décidé de revoir cette grille de statistiques. Grâce à la dynamique de groupe, il a été possible de réaliser une liste exhaustive des types d'interventions des TR. Puis, ces interventions ont été répertoriées en fonction des quatre objectifs à court terme. La nouvelle grille permet de distinguer le sexe, le nombre d'individus rencontrés seuls, en *gangs* ou en groupes formels. Le même exercice a ensuite été effectué pour les problématiques rencontrées. Pour aider les TR dans cette réflexion délicate, les

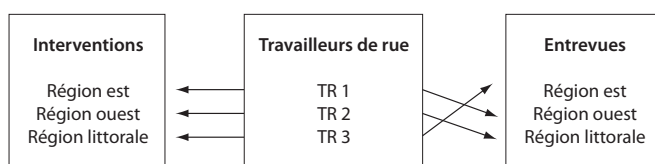
évaluateurs ont proposé d'employer un cadre conceptuel en santé des populations (Dahlgren et Whitehead, 1992). Ainsi, les problématiques touchées par le projet ont été regroupées sous quatre déterminants de la santé : le comportement individuel ; le réseau social, familial et communautaire ; les conditions de vie et de travail ; et les conditions socioéconomiques, culturelles et environnementales.

- Il a été nécessaire de transformer le journal de bord employé par les TR en « portraits d'interventions ». L'objectif d'un tel outil était de permettre aux TR de caractériser leur pratique en fonction des objectifs à court terme, un moyen plus pertinent et utile pour porter un jugement sur l'efficacité du projet. Ce portrait permettait de plus de dégager les facteurs ayant facilité les interventions avec effets, qualifiées de positives par les TR, et les facteurs ayant fait en sorte que d'autres soient négatives, soit sans effets. Pour avoir suffisamment de données sans pour autant alourdir inutilement les tâches administratives, il a été décidé avec les acteurs du projet de limiter la collecte de données à un maximum de deux interventions (positive et négative) par objectif et par semaine. Cette collecte s'est déroulée durant sept mois au cours de l'année 2002.
- Enfin, le dernier outil était l'entrevue individuelle non directive. Il était en effet indispensable d'obtenir le point de vue des jeunes. Plusieurs difficultés inhérentes à la réalisation de ce type de collecte dans ce contexte particulier ont été soulevées dès le début : trouver les jeunes, leur permettre d'être à l'aise pour parler et se confier à un évaluateur externe, etc. Pour supprimer ces obstacles et rester en phase avec l'approche évaluative choisie, il a été décidé que les entrevues seraient réalisées par les TR. D'un point de vue pratique, puisque le territoire était découpé en trois zones sous la responsabilité de trois équipes disposant chacune d'un TR à plein temps, il a été convenu que chacun d'entre eux devait réaliser cinq entrevues. Il était évidemment délicat, d'un point de vue méthodologique et éthique, de demander aux jeunes de parler des interventions du travail de rue à un TR avec lequel ils étaient en contact. Ainsi, pour réduire les risques de biais, il a été proposé que les TR changent de territoire pour effectuer leurs entrevues (figure 15.1). En ce qui concerne le recrutement des jeunes, ces derniers étaient choisis par les TR de leur secteur, puis présentés à celui qui devait mener l'entrevue, le premier quittant les lieux lorsque l'entrevue

commençait. Pour éviter que les jeunes soient mal à l'aise, les entrevues n'ont pas été enregistrées, mais leurs propos ont été pris en note à l'aide d'un canevas de questions préalablement construit avec tous les participants à l'évaluation. Un guide a été produit pour fournir aux TR quelques conseils méthodologiques concernant la tenue d'entrevues et la prise de notes. En tout, 17 entrevues ont été effectuées. Les sujets choisis ont été retenus en fonction de leur accessibilité (Patton, 2002), tout en essayant de respecter un équilibre lié au processus d'intervention des TR avec ces jeunes.

FIGURE 15.1

Stratégie de répartition des entrevues



La stratégie d'analyse

La stratégie d'analyse nécessite une présentation, puisque la démarche inhérente à l'évaluation vise en partie à renforcer les compétences et capacités des intervenants en évaluation de programmes. L'analyse des données qualitatives, qui constituent la majeure partie du corpus de données, a été entièrement réalisée par les TR. Pour cela, des séances de travail ont rassemblé cinq TR à trois reprises. Pour qu'ils soient en mesure d'en faire l'analyse, une réunion de préparation a été organisée, au cours de laquelle un guide d'analyse a été distribué et expliqué. L'organisation d'un exercice de groupe à partir des données déjà disponibles à ce moment-là (trois entrevues) a permis aux participants de mieux intégrer la démarche d'analyse basée sur cinq étapes distinctes : relire les questions d'entrevue, lire les notes des entrevues, discuter des réponses en groupe, catégoriser les réponses et résumer les résultats (Aubel, 2000).

Les données issues des portraits d'intervention ont été analysées par les trois TR impliqués le plus longtemps dans le projet. De façon intuitive, ils ont appliqué une technique recommandée dans certains cas en analyse qualitative. Pour chacun des quatre objectifs à court terme, ils ont réper-

torié les lieux d'interventions, les problématiques touchées, les facteurs ayant facilité ou entravé les interventions. Ces caractéristiques ont été quantifiées et celles dont la fréquence a été la plus importante ont été analysées un peu plus en profondeur que les autres. Il ne s'agit pas d'une analyse quantitative à proprement parler, mais les données chiffrées (leur nullité comme leur surnombre) permettent de repérer certains éléments nécessitant une attention particulière. Une synthèse de plusieurs pages a été produite à la suite de ce processus.

La seule partie de l'évaluation qui a été réalisée par des évaluateurs externes est l'analyse (descriptive) des données quantitatives. Les résultats produits ont cependant été validés par les participants à l'évaluation.

Les résultats

De prime abord, il est essentiel de préciser qu'il faut rester humble et ne pas trop s'attarder aux effets les plus distaux du projet. Autrement dit, compte tenu de l'état d'avancement de l'implantation du projet à Bellechasse et des nombreux changements dans le personnel (ce qui est une caractéristique des projets de « travail de rue »), il ne fallait pas s'attendre à une efficacité optimale dans l'atteinte de l'objectif ultime du projet, soit la responsabilisation et l'autonomie des jeunes. Quelques indices laissent croire que certains jeunes ont eu des relations privilégiées avec les travailleurs de rue, ce qui permet de penser que cet objectif a été atteint pour eux. Les intervenants ont pu développer une expertise et des stratégies particulières pour contribuer à l'émancipation des jeunes. L'objectif qui paraît le moins efficacement atteint, à la lumière des données recueillies, est celui de l'orientation des jeunes vers d'autres ressources. Il est intéressant de relever qu'une recherche très récente sur le travail de rue dans la région de l'Abitibi-Témiscamingue en arrive à la même conclusion (Simard *et al.*, 2002). Très peu de références à diverses ressources sont effectuées, et lorsqu'elles ont lieu, c'est parce qu'il existe un lien de confiance avec les intervenants appartenant aux ressources institutionnelles, phénomène qui a aussi été constaté à Bellechasse. En revanche, les données disponibles à la suite de l'évaluation permettent de conclure que les deux premiers objectifs, celui de dépister les jeunes ciblés par le programme, puis de les informer et de les soutenir en fonction de leurs besoins personnels, ont été atteints. Il s'agit des résultats les plus probants pour le projet de Bellechasse.

Cette évaluation a également permis de révéler certaines stratégies adoptées par les travailleurs de rue pour amorcer des relations et tisser des liens avec les jeunes. Il a aussi été possible de détecter certains facteurs ayant facilité ou entravé les interventions des travailleurs de rue.

Malgré ces résultats concluants, il faut rester très prudent quant à leur validité externe, ou leur généralisation, tant il est notoire que « les manifestations de la vie de rue peuvent varier selon les milieux [...] et qu'il faut demeurer prudent face aux généralisations abusives en ne projetant pas la dynamique sociale d'une ville sur une autre » (Parazelli, 2002, p. 46). Cela étant dit, et revenant sur les recommandations des auteurs québécois s'intéressant au travail de rue et demandant que le discours des jeunes soit mieux entendu, il faut globalement retenir que les jeunes ont exprimé très clairement leur appréciation positive de ce projet. Ils ont explicitement demandé que ce dernier perdure, ce qui traduit un degré de satisfaction important. L'ensemble des résultats de cette évaluation est disponible dans le rapport final.

Discussion

Selon Cousins et Whitmore (1998), trois caractéristiques particulières définissent une évaluation participative, et la qualité de l'implication des acteurs dans le processus permet de déterminer si leur participation est de nature pratique ou émancipatrice : le contrôle du processus évaluatif, la sélection des personnes ressources en tant que participants, et enfin la qualité de la participation de ces derniers. À Bellechasse, pour chacune de ces trois caractéristiques, l'implication des TR a été très importante, sinon totale. Concrètement, le type d'évaluation choisi ainsi que la pertinence de l'utilisation de tel ou tel outil ont été discutés entre les acteurs, et les décisions ont été prises de façon concertée, les TR s'étant vu accorder le dernier mot. Les questions posées lors des entrevues ont été choisies d'un commun accord. Les données ont été recueillies par les TR, et les données qualitatives ont été analysées par eux. Les jeunes interrogés ont été choisis par les TR.

La participation active des TR au développement du plan d'évaluation a permis de trouver une solution originale aux difficultés maintes fois rencontrées et inhérentes à l'évaluation de ce type de projets, ce qui ne fut pas toujours facile. Les évaluateurs ont été obligés de s'adapter au

contexte, de vulgariser certaines approches et techniques de collecte et d'analyse des données, d'être souples dans leurs propositions méthodologiques, d'accepter de partager leurs connaissances et donc leur pouvoir, tout en respectant les critères de qualité et d'éthique en matière d'évaluation. Les mêmes leçons avaient d'ailleurs été tirées d'une expérience similaire menée à Ottawa : « La définition de la qualité a été adaptée afin de prendre en compte le caractère inclusif du processus, ainsi que les enjeux de pouvoir et de contrôle » (Whitmore et McKee, 2001, p. 400, notre traduction). Les TR ont pour leur part dû réfléchir aux stratégies les plus appropriées pour recueillir les données sans pour autant alourdir leurs tâches quotidiennes et tout en respectant leur code d'éthique et les jeunes eux-mêmes. Ils ont dû s'astreindre à effectuer la recherche d'informations auprès des jeunes, et surtout, ce qui témoigne d'un engagement très important de leur part, puisqu'ils n'avaient aucune incitation financière à le faire, s'atteler à la tâche de l'analyse des données qualitatives. Aux dires des TR, leurs collègues des autres régions n'auraient certainement jamais accepté de s'investir à ce point, d'autant plus que leurs réticences à voir leur projet évalué étaient grandes. L'implication des acteurs du projet est une stratégie pragmatique (pour réduire les coûts, les obstacles à l'évaluation) tout autant que politique (utilité des résultats, voir chapitre 9) ou idéologique (soit l'*empowerment* dans les champs de la promotion de la santé et de l'évaluation [Fetterman, 2000 ; Rootman *et al.*, 2001]). Nous n'aborderons pas dans ce chapitre les débats épistémologiques épiques concernant subjectivité et objectivité de la collecte et de l'analyse des données qualitatives en évaluation de programmes ; d'autres l'ont déjà fait (Péladeau et Mercier, 1993). Il est cependant intéressant de noter que les TR sont pleinement conscients des problèmes concernant cette question, puisqu'ils disent eux-mêmes : « Il est important de prendre conscience que le bagage de vie de chacun des travailleurs de l'équipe peut teinter les données recueillies. » Mais le fait qu'ils le soulignent témoigne déjà d'une certaine distanciation. Bref, l'expérience de travail en collaboration a été fructueuse, puisqu'il a été possible de répondre aux questions d'évaluation, mais la stratégie adoptée a-t-elle réellement renforcé le pouvoir d'agir des TR ?

Pour répondre à cette question, nous allons nous éloigner des débats des évaluateurs (Fetterman, 2000 ; Cousins, 2005) et nous tourner vers les experts en service social (Ninacs, 2002). Pour mieux vérifier l'atteinte

de l'objectif d'*empowerment* dans la mise en œuvre de notre évaluation – les réflexions de Cousins et Whitmore (1998) ne pouvant nous aider –, nous utiliserons le cadre d'analyse du processus de l'*empowerment* individuel de Ninacs (2002). D'autres tentatives d'utilisation de ce cadre ont été publiées dans un numéro spécial de la *Revue canadienne d'évaluation de programmes* (Ridde, 2006a). Pour ce chercheur, le pouvoir d'agir individuel correspond à une suite d'étapes opérant, tels quatre fils d'une même corde, sur quatre plans: la participation, les compétences techniques, l'estime de soi, la conscience critique. Le passage par ces étapes, ainsi que leur interaction, permet à un individu de passer d'un état sans (ou avec peu de) pouvoir à un état où il est en mesure d'agir en fonction de ses propres choix. Dans notre cas, l'action englobe à la fois l'évaluation des effets de l'intervention et de son implantation auprès des jeunes. Qu'en est-il de l'*empowerment* des TR lors de la mise en œuvre de cette évaluation ?

En ce qui concerne la participation des TR à l'évaluation, nous croyons avoir donné suffisamment d'informations dans la section de ce chapitre consacrée au processus pour pouvoir affirmer qu'elle a été active et non fictive. À l'instar d'Arnstein, Ninacs propose un continuum allant de l'assistance muette à la participation à la prise de décision. Dans le cas qui nous occupe, il est aisé d'affirmer que la participation des TR s'est située à l'extrême positif de ce continuum. Certains TR ont même précisé que leur participation commune a permis au groupe de TR d'être plus interactif. La cohésion du groupe s'en est aussi trouvée renforcée, nous ont-ils dit, les membres de l'équipe s'étant soudés et les personnes s'étant rapprochées.

À propos des compétences techniques, il faut distinguer celles qui concernent l'évaluation et celles qui sont propres à l'intervention auprès des jeunes. Pour ce qui est de l'évaluation, les TR ont affirmé avoir beaucoup appris: «Ça a été très formateur.» Ils nous ont dit s'être approprié de nombreuses techniques évaluatives. Remarquons que l'intervention des évaluateurs externes ne s'est jamais concrétisée en une formation à l'évaluation à proprement parler. De telles séances n'ont pas été organisées, mais des concepts ont été présentés, des outils créés, des notions d'analyse qualitatives abordées. Ajoutons cependant que l'un des TR de l'équipe avait suivi deux formations en évaluation données par des membres d'une firme d'évaluation. Ce TR nous a confié que sa participation à de telles

séances lui a été très profitable, lui ayant permis de réfléchir au projet et de prendre de la distance vis-à-vis de ses interventions. L'évaluation lui a aussi permis d'améliorer ses compétences techniques en intervention auprès des jeunes. Le fait d'avoir eu à interroger les jeunes sur la pratique des TR et à analyser le contenu des entrevues a permis à ce travailleur de mieux comprendre le sens de leurs interventions: « J'en ai appris beaucoup sur les stratégies d'intervention. » Le fait de créer des catégories pour répertorier leurs actions, pour identifier et relever les facteurs qui les ont facilités et pour entendre les jeunes à propos des stratégies qu'ils pensent être les plus propices pour entrer en contact avec eux sont autant d'éléments qui ont contribué à améliorer les compétences d'intervention des travailleurs de rue. À l'inverse, un TR se demande si cette réflexion n'aurait pas nui à son travail en l'obligeant maintenant à observer ses interventions selon un processus linéaire découpé en étapes et trop structuré. Alors que pendant les réunions de travail habituelles on disséquait des cas individuels, lors des analyses requises pour l'évaluation il fallait faire la synthèse de plusieurs interventions en même temps, mettant ainsi en lumière quelques éléments nouveaux et communs à leurs actions. Le fait d'avoir effectué des analyses qualitatives a aussi contribué à changer « mon regard sur les jeunes », nous a dit un TR, puisqu'il est passé du statut d'observateur à celui d'observé. L'un d'entre eux soutient aussi qu'il a pu réfléchir à ses modes d'intervention, allant même jusqu'à affirmer que son style de pratique a changé. Les TR pensent donc que l'évaluation a aussi été utile en termes d'intervention. Enfin, relevons que les TR se sont approprié le modèle des déterminants de la santé présenté plus haut et qu'ils l'utilisent désormais dans leurs interventions avec les jeunes (Ridde, 2007).

Plusieurs éléments nous font affirmer que l'estime personnelle des TR s'est trouvée renforcée: « J'en ressors grandi. » L'évaluation à Bellechasse semble avoir eu un effet bien réel sur les deux éléments fondamentaux de l'estime de soi selon Ninacs. Premièrement, elle a permis aux TR de percevoir qu'ils possédaient des compétences réelles qui avaient des effets indubitables sur les jeunes: « [J'ai vu] que ma simple présence avait des impacts. » Ils s'en sont dits très fiers. Deuxièmement, ils ont pris conscience du fait que leurs compétences étaient reconnues par les autres, en l'occurrence les jeunes participants à leur projet. Enfin, il faut bien reconnaître que l'évaluation n'a pas eu d'effets sur le dernier élément de l'*empowerment* individuel, soit la conscience critique collective, sociale ou politique. La

raison en est simple : les TR sont des intervenants socialement engagés ; ils se définissent comme des agents de changement social et leur implication dans de tels projets relève plus souvent d'un engagement politique que d'une profession. Il s'agit « d'un mode de vie », de « donner du sens à une job », nous ont-ils dit. Cela nous paraît d'autant plus vrai que ceux et celles qui ne peuvent concevoir ce métier selon ces valeurs particulières ne peuvent supporter bien longtemps les conditions difficiles de travail et de rémunération, et quittent rapidement les projets. Ce problème est d'ailleurs récurrent dans ce type d'interventions de rue, et les organismes communautaires s'interrogent beaucoup sur les stratégies à déployer pour contrer ce fort taux de roulement du personnel.

Conclusion

Dans le contexte actuel du rapprochement des services sociaux et de santé des sociétés, les projets de travail de rue auront dans les prochaines années un rôle important à jouer pour répondre aux besoins des jeunes ne désirant pas ou ne pouvant pas s'orienter vers les ressources traditionnelles. Mais ces projets sont très difficiles à évaluer, d'abord parce que leurs acteurs et leurs bénéficiaires ne sont pas encore ouverts à cette forme de reddition de comptes, et ensuite parce que les devis traditionnels d'évaluation ne sont pas en phase avec les valeurs et les interventions de ce type de projets. Dans ce chapitre, nous avons cherché à décrire et à analyser un processus évaluatif au cours duquel la participation des parties prenantes de l'évaluation était basée sur des objectifs d'émancipation et de renforcement du pouvoir d'agir. Nous aurions pu faire plus en impliquant les jeunes (Whitmore et McKee, 2001), mais compte tenu du temps dont nous disposions et des ressources imparties à l'évaluation, le pragmatisme nous a fait pencher pour la solution présentée dans ce chapitre. Le prochain défi évaluatif sera de faire participer les bénéficiaires des projets, et non plus les intervenants. Ici, l'analyse de ce processus nous permet de croire, d'une part, que cette stratégie est la plus susceptible d'être acceptée par les acteurs et de produire des données probantes pour porter un jugement sur l'efficacité des interventions, et, d'autre part, que l'évaluation relatée dans ce cas a réussi, du moins en partie, à renforcer le pouvoir d'agir des TR.

16

Une évaluation des besoins en formation d'intervenants en prévention

François Chagnon, Janie Houle et Isabelle Marcoux

Le contexte

Les promoteurs de formation ont la responsabilité de rendre compte de la pertinence et des effets de leur programme aux bailleurs de fonds. Dans le cas d'une formation en prévention du suicide, cette responsabilité pose des défis importants, car les crises suicidaires sont des événements peu fréquents et il est difficilement envisageable, sur le plan éthique, d'évaluer des interventions portant sur des situations réelles. Comment, dès lors, élaborer un devis d'évaluation permettant de juger de l'efficacité de telles formations? Ce chapitre présente le processus et les résultats d'une évaluation des effets d'un programme de formation en prévention du suicide chez les jeunes qui a été réalisée à l'aide d'une stratégie d'évaluation fondée sur l'approche par compétences.

Au Québec, le tiers des décès observés chez les jeunes de 15 à 19 ans sont la conséquence de suicides, ce qui représente la première cause de mortalité pour les personnes de ce groupe d'âge (Institut de la statistique du Québec, 2003). Le plus souvent, le risque de suicide chez les jeunes qui se sont enlevé la vie n'avait pas été reconnu par les professionnels; aussi

n'avaient-ils pas reçu l'aide dont ils avaient besoin (Brent, 1995). La capacité d'identifier les jeunes à risque et celle de les adresser rapidement à des professionnels compétents sont donc des éléments déterminants de la prévention du suicide (Shaffer et Craft, 1999). Il s'avère donc crucial de mettre en œuvre les stratégies les plus efficaces afin d'améliorer les compétences des intervenants à dépister et à diriger les jeunes à risque de suicide vers les ressources appropriées.

Les enseignants et les intervenants sociaux et communautaires qui œuvrent auprès des clientèles jeunesse occupent des positions privilégiées pour dépister et diriger les jeunes à risque de suicide vers les bonnes ressources. La formation de ces acteurs de première ligne est donc considérée comme une stratégie importante et prometteuse de prévention du suicide chez les jeunes (Gould et Kramer, 2001; Shaffer *et al.*, 2001). Depuis les années 1980, de nombreuses formations en intervention auprès des jeunes suicidaires ont été conçues et dispensées un peu partout au Canada et aux États-Unis. Les similitudes qu'on retrouve dans leurs contenus suggèrent que la logique d'action de ces formations repose sur la théorie suivante : par une activité de formation visant l'acquisition d'attitudes favorables à la prévention du suicide, de connaissances de pointe sur les facteurs de risque ainsi que d'habiletés d'intervention spécifiques, il est possible d'améliorer le dépistage des jeunes à risque, de faciliter leur envoi vers les ressources adéquates et, ainsi, de mieux prévenir le suicide.

De telles formations, si elles semblent fondées, sur le plan théorique, exigent cependant un investissement important en temps et en argent de la part des milieux d'intervention, et cela dans un contexte où les ressources disponibles pour la prévention du suicide sont limitées. Dans le but d'améliorer les efforts en ce sens, il apparaît crucial de vérifier la pertinence du modèle théorique de ces formations ainsi que leur efficacité réelle et, le cas échéant, de déterminer les correctifs pouvant permettre d'augmenter leur capacité d'atteindre les objectifs visés.

Malgré la profusion de formations disponibles, peu de recherches ont examiné leur efficacité, et celles qui l'ont fait présentent des lacunes importantes sur le plan méthodologique. Entre autres, certaines se sont limitées à évaluer l'acquisition des attitudes et des connaissances sans égard au développement des habiletés d'intervention (Davidson et Range, 1999; Stuart *et al.*, 2003). Les études qui ont mesuré le savoir-faire des intervenants se sont heurtées aux difficultés inhérentes à une telle entre-

prise. En effet, la méthode idéale consisterait à observer les participants dans leur milieu naturel au moment où ils interviennent auprès d'un jeune suicidaire. Toutefois, en raison de la rareté des comportements suicidaires et des coûts importants associés à cette stratégie, cette méthode n'a jamais été utilisée. Deux études ont opté pour une méthode intermédiaire en évaluant les habiletés d'intervention des intervenants à l'aide de simulations (Tierney, 1994; Haw et Andres, 1998). Les participants devaient interagir avec un comédien qui simulait une crise suicidaire et leurs habiletés étaient évaluées par une tierce personne à l'aide d'une grille d'observation. Cette méthode a l'avantage d'être celle qui ressemble le plus à une situation réelle, ce qui en accroît la validité; son coût est cependant très élevé.

Parmi les autres méthodes d'évaluation des habiletés d'intervention, on trouve des vignettes imprimées mettant en scène des personnes suicidaires et pour lesquelles les participants doivent décrire la conduite qu'ils adopteraient dans ces situations fictives. Dans leur étude, Capp, Deane et Lambert (2001) ont adopté cette méthode pour évaluer l'intention des participants d'adopter un comportement donné face à un jeune suicidaire à partir de questions à choix de réponses.

En plus de présenter les limites inhérentes aux diverses méthodes utilisées pour mesurer les habiletés d'intervention, une seule étude a vérifié si les compétences acquises au cours des formations se maintenaient dans le temps. Les résultats de cette étude ont révélé que les gains observés après une formation de pairs aidants demeuraient significatifs trois mois après la formation (Stuart *et al.*, 2003).

En somme, la formation des intervenants est une stratégie de prévention du suicide très répandue, mais la mesure de ses effets quant à l'acquisition d'habiletés d'intervention pose des défis méthodologiques importants, et jusqu'à maintenant, il existe peu de preuves du maintien de ces compétences plusieurs mois après la fin de ces formations.

L'évaluation présentée ici est le fruit d'une collaboration entre le Centre de recherche et d'intervention sur le suicide et l'euthanasie (CRISE) et Suicide Action Montréal (SAM), le plus important centre de prévention du suicide (CPS) au Québec. Depuis des dizaines d'années, les professionnels de SAM forment les intervenants de la communauté à la prévention du suicide. Ils détiennent une bonne crédibilité dans la région de Montréal et peuvent offrir leurs formations gratuitement grâce au soutien financier

de quelques bailleurs de fonds. Cependant, malgré leurs nombreuses années d'expertise, leur formation orientée sur la clientèle jeunesse n'avait jamais fait l'objet d'une évaluation visant à en apprécier la qualité et l'efficacité. L'enjeu est ici de taille, non seulement en raison de l'importance du suicide chez les jeunes au Québec et du potentiel de généralisation des résultats aux autres CPS de la province, mais aussi du fait que le financement ultérieur de la formation par les bailleurs de fonds sera tributaire des résultats de cette évaluation.

Le protocole d'évaluation

L'objectif et les questions d'évaluation

L'objectif principal poursuivi dans le cadre de cette évaluation est de vérifier l'effet d'une formation spécialisée quant à l'acquisition et au maintien de compétences en prévention du suicide chez les jeunes. Cette évaluation voulait ainsi répondre à quatre questions d'évaluation prioritaires : 1) le programme de formation parvient-il réellement à améliorer les attitudes, les connaissances et les habiletés des intervenants ; 2) dans la mesure où de tels gains sont observés, ces habiletés se maintiennent-elles dans le temps ; 3) quels sont les contenus de la formation les plus retenus par les participants et, en contrepartie, quels sont les éléments les moins retenus ; et 4) quelles sont les avenues prioritaires pour améliorer cette formation ?

La méthode

L'évaluation a comporté deux volets successifs. Dans un premier temps, nous avons procédé à une analyse théorique et clinique des contenus de la formation afin d'examiner en profondeur la pertinence des contenus pédagogiques dispensés aux intervenants. Cette stratégie avait pour but d'améliorer le plus possible la formation avant même d'en évaluer les effets, de faciliter la création d'outils de mesure parfaitement adaptés aux contenus et de favoriser une utilisation maximale des résultats de l'évaluation en permettant de réaliser une analyse détaillée des effets de la formation.

Dans un deuxième temps, nous avons procédé à la phase d'expérimentation en comparant des intervenants exposés à la formation (groupe

expérimental) à des intervenants ne l'ayant pas été (groupe témoin). Un devis à mesures répétées a été utilisé. Les participants à la condition expérimentale ont été évalués à trois reprises : avant la formation, immédiatement après et six mois plus tard. Les deux premières évaluations ont été effectuées dans le lieu où était dispensée la formation, et la dernière évaluation, dans leur lieu de travail. Les participants du groupe témoin étaient quant à eux évalués dans leur milieu de travail à deux reprises, à trois semaines d'intervalle. En tout, 78 intervenants œuvrant dans des établissements scolaires et des organismes communautaires ou institutionnels desservant la clientèle jeunesse de la région de Montréal ont été recrutés. Ils ont été assignés à la condition expérimentale ou témoin en fonction de la date de leur inscription à la formation. Il faut noter que les participants du groupe témoin ont bénéficié gratuitement de la formation une fois l'étude terminée. Pour des raisons éthiques visant notamment à assurer l'équité entre tous les participants, il s'avérait en effet important d'offrir le même service à tous dans des délais raisonnables.

Les mesures

Un questionnaire auto-administré a été conçu spécifiquement pour cette étude, puis prétesté auprès de 14 intervenants. Conformément à l'approche par compétences de Rycus et Hugues (2000), trois éléments ont été mesurés : les *attitudes* à l'égard de l'intervention auprès des personnes suicidaires, les connaissances et les habiletés d'intervention.

Les attitudes. Un instrument de mesure standardisé, le *Suicide Intervention Questionnaire* (SIQ) de Tierney (1988), a été utilisé pour mesurer les attitudes. Cet instrument compte 20 énoncés à propos desquels le répondant doit indiquer son niveau d'accord sur une échelle de type Likert à 5 points, allant de (5) « tout à fait en accord » à (1) « tout à fait en désaccord ». Le SIQ possède de bonnes qualités psychométriques avec un coefficient (Pearson) de fidélité interne de 0,79 et une fidélité test-retest de 0,82.

Les connaissances. Le questionnaire utilisé pour mesurer les connaissances a été élaboré spécifiquement pour cette étude suivant le modèle de Rycus et Hugues (voir tableau 16.1). Il comporte 12 questions ouvertes permettant de mesurer l'acquisition des connaissances dans six domaines : 1) facteurs de risque et de protection du suicide ; 2) éléments essentiels

d'une intervention de crise; 3) lois en matière de confidentialité avec des mineurs; 4) actions à mettre en place à la suite d'une tentative de suicide; 5) étapes du processus de deuil; 6) responsabilités confiées aux intervenants. Un point est attribué par question, de sorte que le score au questionnaire mesurant les connaissances peut varier de 0 à 12.

Les habiletés d'intervention. Les habiletés essentielles que les intervenants doivent maîtriser sont mesurées à l'aide de sept vignettes cliniques, dont cinq imprimées et deux enregistrées sur vidéocassette. Le répondant doit préciser ce qu'il convient de faire dans chacune des situations critiques décrites sur les vignettes. Ces simulations mettent en scène un intervenant confronté à diverses situations: a) il ne se sent pas à l'aise pour intervenir auprès d'un jeune suicidaire; b) il doit mettre en place des actions auprès de jeunes qui viennent d'apprendre qu'un de leurs amis a tenté de se suicider; c) il doit faire un rapport d'événement au sujet d'un jeune qui fréquentait son organisme et qui a fait une tentative de suicide; d) il doit repérer les jeunes devenus vulnérables à la suite du suicide d'un autre jeune; e) il doit faire une intervention de crise auprès d'un jeune suicidaire.

Les deux vignettes vidéo sont d'une durée approximative de trois minutes chacune, et leur contenu a été élaboré conjointement par un membre de l'équipe de recherche et une formatrice, qui se sont basés sur une analyse critique de la littérature scientifique et sur l'expérience clinique d'un groupe d'experts. Pour chacune des vignettes, le répondant doit préciser s'il juge que le jeune est à risque de suicide; expliquer les éléments sur lesquels il base son évaluation du risque; distinguer les facteurs de risque des signes précurseurs; évaluer l'urgence de la situation; préciser les éléments qui lui permettent de statuer sur l'urgence de la situation; déterminer quelles informations complémentaires il est essentiel d'obtenir auprès du jeune pour guider son intervention et préciser les actions qu'il poserait à court terme dans une telle situation. Enfin, deux questions à choix de réponses portant sur la directivité et le niveau de confiance nécessaire pour intervenir auprès d'un jeune suicidaire complètent la mesure des habiletés d'intervention.

La description du processus d'évaluation

Pour l'évaluation de la formation en prévention du suicide chez les jeunes, une approche participative a été adoptée afin d'assurer une bonne crédi-

bilité des résultats de l'étude et de maximiser leur utilisation par l'organisme promoteur, SAM. Dans cette perspective, deux comités de travail ont été créés : un comité de coordination et un comité d'experts.

Le comité de coordination a été constitué afin d'assurer la bonne marche du projet et de favoriser la participation des promoteurs de la formation à toutes les étapes de la démarche d'évaluation. Outre les membres de l'équipe de recherche, le comité de coordination regroupait un gestionnaire clinicien responsable de la formation jeunesse dispensée par SAM, deux formatrices professionnelles ainsi que, sur une base *ad hoc*, le directeur de l'organisme. Tout au cours de l'exercice, les membres du comité de coordination ont collaboré de façon étroite afin de concevoir les outils de mesure, de recruter les participants, d'analyser et d'interpréter les résultats de l'évaluation et de s'assurer que celle-ci atteigne les objectifs fixés.

Le comité d'experts cliniciens avait les différents mandats suivants : définir les compétences attendues des intervenants de première ligne en prévention du suicide ; analyser la pertinence théorique et clinique des contenus abordés et leur concordance avec les compétences visées par la formation ; soutenir l'élaboration des outils de mesure conçus spécifiquement pour cette évaluation. Ce comité regroupait huit experts issus des milieux pédopsychiatrique, scolaire et communautaire.

Pour le processus d'évaluation, nous avons adopté la démarche proposée par le modèle de compétences de Rycus et Hugues (2000) mis au point dans le domaine des services à l'enfance, qui est fréquemment utilisé pour déterminer les compétences attendues des intervenants dans des domaines d'intervention spécifiques (voir tableau 16.1).

TABLEAU 16.1

Les étapes du modèle de compétences de Rycus et Hugues

1	Définition de la population à former
2	Analyse des responsabilités et tâches des intervenants
3	Spécification des standards de qualité d'une bonne pratique
4	Délimitation des responsabilités des intervenants
5	Analyse des compétences attendues dans l'exercice de ces responsabilités
6	Détermination d'indicateurs permettant d'apprécier l'application de ces connaissances

La première étape du processus d'évaluation fut de décrire les compétences attendues des intervenants de première ligne en matière de prévention du suicide quant au savoir, au savoir-être et au savoir-faire. Cette

étape était un préalable à la détermination des indicateurs pertinents et pour l'élaboration des instruments de mesure. Conformément au cadre d'analyse développé par Rycus, le comité de coordination s'est engagé dans un intense processus de collaboration avec le groupe d'experts afin d'identifier les compétences devant être maîtrisées par les intervenants de première ligne et de vérifier la concordance entre ces compétences et le contenu de la formation dispensée par l'organisme promoteur. Au cours de cette étape déterminante, les partenaires ont profité des fruits des recherches d'un groupe de travail multidisciplinaire qui, sous la responsabilité de l'Association des centres jeunesse du Québec, avait déjà examiné la question de l'intervention en matière de prévention du suicide chez les jeunes en centre jeunesse et élaboré un protocole d'intervention balisant les rôles et les responsabilités des intervenants de première et de deuxième ligne.

Ce processus de travail s'est déroulé sur une période de six mois et a mené à un consensus sur huit compétences essentielles à acquérir pour qu'un intervenant puisse faire correctement son travail de dépistage et qu'il dirige les jeunes suicidaires vers les bonnes ressources (voir tableau 16.2). Par la suite, ces huit compétences générales ont été subdivisées en 59 sous-énoncés de compétences spécifiques à acquérir.

TABLEAU 16.2

Les compétences attendues des intervenants de première ligne en matière de prévention du suicide chez les jeunes

1	Connaître et détecter les facteurs de risque et les facteurs de protection
2	Connaître et détecter les signes de détresse psychologique, les indices de trouble mental et les signes précurseurs du suicide
3	Reconnaître la présence d'indices d'urgence et de risque élevé
4	Connaître les responsabilités des intervenants de première et de deuxième ligne
5	Connaître les principales règles de confidentialité
6	Connaître le mode d'intervention en contexte de crise suicidaire et savoir l'appliquer
7	Savoir intervenir après une tentative de suicide ou un suicide
8	Savoir assurer la prise en charge des interventions auprès d'un jeune en crise suicidaire

Cette stratégie d'analyse qualitative s'est avérée extrêmement profitable, puisqu'elle a permis non seulement de valider le contenu de la formation dispensée par l'organisme promoteur, mais également de le bonifier. En outre, cette démarche de consensus entre experts de différentes disciplines aura contribué à accroître la crédibilité de la formation et à établir une

alliance entre ces groupes de professionnels. Enfin, cette analyse de contenu a permis de mettre en place les bases conceptuelles nécessaires au développement des instruments de mesure.

Ainsi, l'équipe de coordination a produit un questionnaire permettant de mesurer chacune des compétences essentielles et qui a ensuite été soumis à l'approbation du groupe d'experts pour en assurer la validité. Une grille de codification a également été élaborée en collaboration avec le groupe d'experts de manière à déterminer, pour chacune des questions ouvertes, les éléments de réponse recherchés. Le résultat de chacune des questions varie en fonction du nombre de bons éléments mentionnés par le participant. Chacune des questions a été normalisée en utilisant une constante afin d'homogénéiser leur poids relatif.

Les résultats de l'évaluation

L'échantillon final comptait 71 sujets, soit 46 femmes (65 %) et 25 hommes (35 %) répartis en deux groupes. Le premier réunissait 43 participants ayant suivi la formation en prévention du suicide chez les jeunes (groupe expérimental) ; le deuxième comptait 28 participants n'ayant pas suivi la formation (groupe témoin). Tel que précisé plus haut, les participants ont été assignés au groupe expérimental ou témoin en fonction de leur choix concernant la date où ils pouvaient suivre la formation. Nous avons exclu cinq sujets de l'échantillon initial parce qu'ils avaient fourni des données incomplètes.

Les participants des deux groupes se sont révélés comparables quant à leur niveau de scolarité, leur environnement de travail, leur nombre d'années d'expérience en intervention auprès de jeunes et auprès de jeunes en situation de crise suicidaire. Près du tiers des participants ont rapporté avoir déjà reçu une formation en prévention du suicide, en moyenne 4,5 ans avant le début de notre étude.

Avant la formation

Les participants des deux groupes étaient comparables sur le plan des attitudes et sur celui des habiletés, mais une différence modeste, quoique statistiquement significative, a été observée en ce qui concerne les connaissances. En effet, lors de la première procédure de mesure, le groupe

expérimental a obtenu un score légèrement plus élevé que le groupe témoin (groupe expérimental : moyenne = 6,47 ; ET = 1,38 ; groupe témoin : moyenne = 5,38 ; ET = 1,55 ; $F_{1,70} = 9,632$; $p = 0,003$).

Immédiatement après la formation

Après avoir suivi la formation, les participants du groupe expérimental se sont significativement améliorés sur le plan des connaissances, des attitudes et des habiletés d'intervention par rapport aux sujets témoins (Willk's Lambda 0,403 ; $F_{3,64} = 31,583$, $p < 0,001$). Ces résultats demeurent statistiquement significatifs, même lorsqu'on contrôle l'effet de la différence observée avant la formation sur le plan des connaissances. Afin de faciliter l'interprétation des résultats, le tableau 16.3 présente les scores pondérés en pourcentage.

TABLEAU 16.3

Résultats obtenus par les participants des groupes expérimental et témoin aux procédures de mesure des compétences, aux trois temps de mesure (N = 71 %)

	Prétest	Après la formation	Six mois après la formation
Connaissances			
Groupe expérimental	54*	69*	61*
Groupe témoin	45	46	–
Attitudes			
Groupe expérimental	76	82*	83
Groupe témoin	78	77	–
Habiletés			
Groupe expérimental	52	67*	62*
Groupe témoin	48	50	–

* $p = 0,003$

Six mois après la formation

Dix participants (23,3 %) du groupe expérimental n'ont pas participé au troisième temps de mesure, mais ils sont comparables aux participants qui ont terminé le processus d'évaluation. Six mois après la formation, les connaissances ($F_{1,32} = 13,465$, $p < 0,001$) et les habiletés ($F_{1,32} = 18,290$, $p < 0,001$) ont significativement diminué par rapport à leur niveau

immédiatement après la formation, alors que les attitudes sont demeurées stables ($F [1,32] = 0,420, p = 0,522$).

Des comparaisons *a posteriori* ont toutefois révélé que, malgré cette diminution sur le plan des apprentissages, le niveau de compétences six mois après la formation demeurait significativement plus élevé que celui d'avant la formation, et cela autant en ce qui concernait les connaissances ($F [1,32] = 12,154, p < 0,001$) que les attitudes ($F [1,32] = 30,194, p < 0,001$) et les habiletés ($F [1,32] = 37,742, p < 0,001$).

Les analyses par item

Afin de contribuer à l'amélioration des contenus de formation et de fournir les données les plus précises possibles aux promoteurs, une seconde vague d'analyse a été réalisée pour chaque item des questionnaires à la lumière des résultats obtenus. Ceux-ci ont été présentés par blocs de contenus sous forme de tableaux vulgarisés utilisant un code de trois couleurs. Un item obtenait un code vert si le score moyen après la formation était supérieur ou égal à 75 %, un code jaune était utilisé pour les résultats variant de 50 à 75 % et un code rouge distinguait les items dont le score était inférieur à 50 %. De cette manière, les formateurs pouvaient savoir quels étaient les contenus sur lesquels ils devaient mettre davantage l'accent pour améliorer leur performance.

Discussion

Le principal objectif de cette évaluation était de vérifier l'effet d'une formation spécialisée en intervention auprès des jeunes en crise suicidaire sur les compétences des intervenants. Dans cette étude participative à la fois sommative et formative, on a donc adopté une méthode novatrice pour déterminer, sur la base du modèle de Rycus et Hugues (2000), des indicateurs d'appréciation des formations à partir des responsabilités et des compétences attendues chez les intervenants.

De manière générale, les résultats de l'étude montrent que cette formation limitée dans le temps a amélioré significativement les connaissances, les attitudes et les habiletés des intervenants. En outre, malgré une diminution significative des acquis après six mois, le niveau de compétence des participants à la formation demeure tout de même plus élevé

que ce qu'il était avant qu'ils la suivent. Cet aspect est crucial compte tenu de l'investissement important en ressources humaines et financières que nécessitent de telles formations.

Une analyse plus fine des résultats a néanmoins révélé que même si les gains étaient statistiquement significatifs, ils demeuraient cependant modestes. En effet, l'amélioration maximale est de l'ordre de 15 %, un succès qui peut sembler bien relatif. Toutefois, il convient de se demander quel peut raisonnablement être le niveau d'amélioration des compétences d'un groupe d'intervenants dans le cadre d'une formation de trois jours, si on le compare par exemple à celui qui peut être observé au terme d'une année de formation universitaire. En outre, si on transpose les résultats de l'évaluation sous forme de « bulletin final », celui-ci indique qu'au sortir de la formation, les participants atteignent un peu plus que la note de passage de 60 %. Cette performance est-elle acceptable? Quels résultats serions-nous en droit d'attendre d'une telle formation? Ces questions méritent réflexion.

Dans cette évaluation, nous avons employé une stratégie d'analyse plus poussée que la stratégie classique consistant à comparer des moyennes et des écarts-types. Nous avons procédé à une analyse en profondeur des contenus théoriques et des résultats obtenus pour chacune des compétences visées, ce qui a permis aux partenaires de l'évaluation de déterminer les contenus qui étaient les moins bien assimilés par les participants et donc de revoir la manière dont ils étaient présentés dans le cadre de la formation. Cette analyse a également mis en évidence le fait que certains contenus avaient une importance plus déterminante que d'autres et qu'ils méritaient donc des efforts plus soutenus sur le plan pédagogique.

Ainsi, l'analyse du modèle de compétence a montré que, malgré les améliorations significatives observées sur le plan statistique, les intervenants ne maîtrisaient que partiellement les connaissances et les habiletés jugées importantes à acquérir dans le contexte d'une formation des intervenants en matière de prévention du suicide chez les jeunes, puisque des éléments très importants ne semblaient pas avoir été acquis. Ces conclusions ont eu pour effet de stimuler, au sein de l'organisme, une remise en question des approches pédagogiques utilisées et des contenus abordés dans la formation. Après l'évaluation, l'organisme a mis en place un comité chargé de revoir la pertinence d'une formation ponctuelle et d'examiner l'hypothèse d'une formation continue pour atteindre les

objectifs d'apprentissage visés, c'est-à-dire générer des apprentissages qui se maintiendraient dans le temps. D'autre part, profitant des avancées de l'analyse des compétences réalisée dans la première partie de l'évaluation, l'organisme a révisé les domaines d'apprentissage visés par sa formation afin de former les intervenants de façon plus intense mais sur un spectre moins large de compétences jugées prioritaires.

Deux défis principaux se sont posés dans le cadre de cette évaluation. D'une part, les attentes exprimées par les bailleurs de fonds ont créé une pression certaine pour les promoteurs de la formation, qui se trouvaient ainsi dans la position où ils devaient démontrer l'efficacité de leur programme de formation. Cette situation initiale était loin d'être idéale pour sceller l'alliance entre l'équipe de recherche et les promoteurs du programme. Bien au contraire, un tel contexte n'a fait qu'ajouter au stress d'être évalué et a obligé l'équipe de recherche et les promoteurs du programme à demeurer en contact étroit pour établir les enjeux de l'évaluation et créer le climat de confiance essentiel au succès de cette entreprise.

À ce contexte particulier s'ajoutait aussi le défi, sur le plan méthodologique, posé par la mesure du concept de compétence. En effet, ce concept devient particulièrement complexe dans un contexte où la mesure de l'effet (ici, l'efficacité de la formation en prévention du suicide) ne peut être réalisée en situation réelle, compte tenu de la rareté des événements suicidaires et du fait qu'il n'existe pas de mesure valide des habiletés des intervenants en prévention du suicide. L'utilisation de vignettes écrites et vidéo s'est avérée un compromis acceptable, mais leur caractère fictif et la nécessité de répondre par écrit plutôt que verbalement nous ont incités à la prudence quant à la généralisation de ces résultats à des situations réelles. Enfin, la validité des résultats de cette évaluation est également limitée en raison de la taille restreinte de l'échantillon et de la répartition non aléatoire des participants dans les deux groupes.

Conclusion

Les résultats de cette évaluation confirment l'hypothèse que des formations à l'intervention peuvent améliorer les compétences des intervenants de première ligne auprès des jeunes à risque de suicide et que les effets de telles formations peuvent se maintenir plusieurs mois après la formation.

Le choix d'une approche d'évaluation participative, bien qu'exigeant en termes d'investissement de la part des partenaires, incluant les membres de l'équipe d'évaluation, s'est néanmoins avéré fructueux dans ce contexte et a permis une forte appropriation des résultats de l'évaluation par les promoteurs. Cette stratégie a également produit un matériel riche et susceptible de faciliter une révision du contenu et de la pédagogie même de la formation.

17

Une évaluation des processus d'un programme de protection de la jeunesse

Christian Dagenais et Didier Dupont

Le contexte

L'évaluation décrite dans ce chapitre découle de l'initiative d'un groupe d'intervenants et de gestionnaires du Centre jeunesse de Montréal – Institut universitaire (CJM-IU) qui souhaitent évaluer l'efficacité des interventions de l'équipe d'Intervention rapide et intensive rattachée à l'accueil (désormais IRI-Accueil) du Directeur de la protection de la jeunesse (DPJ). Au mois d'avril 2002, le CJM-IU déposait au Centre d'excellence sur la protection et le bien-être des enfants une demande de financement pour l'évaluation du programme d'intervention offert par cette équipe.

L'intervention à évaluer est mise en œuvre lorsqu'une situation de crise familiale signalée au Directeur de la protection de la jeunesse (DPJ) pourrait nécessiter le placement d'un jeune en urgence. Un intervenant de l'équipe IRI-Accueil est alors immédiatement dépêché sur les lieux pour tenter de résorber la crise, pour obtenir des informations complémentaires sur la situation de ce jeune et de sa famille, et pour consolider le processus de référence vers les ressources de la communauté afin d'éviter de retenir

inutilement ce signalement. L'intervention de l'équipe IRI-Accueil s'apparente au modèle américain *Homebuilders* et en respecte les principes directeurs. En tout, six intervenants et intervenantes forment l'équipe.

Le modèle d'intervention

Le programme IRI-Accueil du CJM-IU s'adresse principalement à une clientèle âgée de 6 à 17 ans signalée pour troubles de comportement et dont la famille est dans une situation d'urgence et de crise. Ce programme poursuit essentiellement trois objectifs : résorber la crise, éviter de procéder à un placement en urgence et consolider le processus de référence vers les ressources de la communauté afin d'éviter de retenir un signalement.

L'intervention de l'équipe IRI-Accueil respecte les principes directeurs du modèle *Homebuilders*, qui sont les suivants : rapidité, intensité, souplesse, déroulement des activités dans l'environnement familial, durée limitée à huit jours, implication de tous les membres de la famille, concertation avec les partenaires et enfin offre de soutien concret (Nelson *et al.*, 1990 ; Dagenais et Bouchard, 1996).

L'élaboration de programmes spécifiquement destinés à maintenir l'enfant dans un milieu familial plus sécuritaire suscite un attrait évident. De nombreux programmes de ce type ont été conçus et largement disséminés depuis le milieu des années 1970 (Kinney *et al.*, 1990 ; Kinney *et al.*, 1991 ; Garant, 1992). Au Centre jeunesse de Montréal, plusieurs initiatives ont été mises en œuvre afin d'éviter le recours au placement lorsqu'il n'est pas indispensable ni nécessaire pour le bien-être de l'enfant (Gouvernement du Québec, 1991 ; Garant, 1992). Ainsi, les projets Ado-RTS-Urgences sociales, Crise apprivoisée, PRIME, Jessie, Négligence 0-5 ans, Urgence ouest/enfance et Intervention psychosociale intensive (Archambault et Des Groseillers, 1997) ont tour à tour permis de tirer des leçons aujourd'hui mises à profit dans l'intervention de l'équipe IRI-Accueil. L'hypothèse à la base du volet IRI-Accueil découle de ces expériences et se formule comme suit : on peut éviter le placement en urgence ou la rétention inutile d'un signalement en offrant sur-le-champ et à domicile un soutien intensif à la famille pour l'aider à faire face à la situation de crise et donc écarter les risques de compromission de la sécurité et du développement de l'enfant. Selon les concepteurs de l'intervention IRI-Accueil, le fait d'offrir une intervention appropriée à la famille permettrait également, le cas

échéant, de prévenir un placement à plus long terme ou la mise en place de mesures de suivi plus lourdes.

Les acteurs et enjeux de l'évaluation

Six intervenants forment l'équipe IRI-Accueil, mais il importe de souligner ici qu'au moment de la collecte de données auprès des intervenants du programme, quatre d'entre eux venaient tout juste de se joindre à l'équipe. Aucune de ces personnes n'avait d'expérience préalable en matière d'intervention rapide et intensive. Elles étaient donc en plein processus d'adaptation à un nouveau mode de pratique.

Les évaluateurs se sont joints à des membres du personnel du CJM-IU pour former un comité de suivi chargé de formuler les questions d'évaluation, de définir les cibles d'évaluation, de procéder au choix et à l'élaboration des outils de collecte des données. Ce comité s'est ensuite élargi au moment du démarrage de l'évaluation. Nous reviendrons sur le rôle central qu'a joué ce comité tout au long du projet dans la section de ce chapitre portant sur le processus d'évaluation.

Le plan d'évaluation

À l'origine, le projet des demandeurs visait à mesurer les effets du programme en comparant les familles qui bénéficiaient des services du programme IRI avec des familles aux prises avec le même genre de problèmes, mais à qui on offrait les services courants du CJM-IU. Dès le début du recrutement des familles, en octobre 2002, nous avons constaté qu'il serait impossible de former le groupe de comparaison planifié. Cela s'explique par le fait qu'entre le moment où nous avons soumis la proposition pour le financement et le début de la recherche, le nombre d'intervenants dans l'équipe est passé de quatre à six. Comme le groupe de comparaison devait être composé de familles admissibles au programme mais non desservies faute de place, l'augmentation du nombre d'intervenants a permis de les desservir, éliminant ainsi la possibilité de former ce groupe. Après réflexions et plusieurs discussions, il fut proposé de mesurer les effets du programme en constituant deux groupes à partir des usagers ayant bénéficié des services du programme IRI-Accueil. Les groupes constitués étaient formés, d'une part, des usagers pour qui l'intervention

s'est avérée terminale et, d'autre part, de ceux pour qui elle a donné lieu à d'autres types de services du CJM-IU ou à un signalement. Cette proposition a exigé une reconfiguration complète du devis d'évaluation prévu à l'origine, reconfiguration qui a été l'occasion d'examiner la mise en œuvre du programme sous deux angles : le processus d'adressage des cas au programme et les processus de mise en œuvre des principes directeurs du programme par les intervenants. Comme nous le verrons, le nouveau devis porte un regard sur les critères utilisés par les intervenants de l'accueil du DPJ pour orienter les familles vers le programme. Il comporte aussi un examen attentif des pratiques des intervenants du programme, ce qui permet de mieux expliquer et comprendre comment l'application du modèle sous-jacent proposé mène ou non aux effets escomptés. Globalement, le devis permet de mieux cerner les clientèles pour qui le programme mène à une intervention terminale et les caractéristiques des interventions en lien avec la fermeture du dossier. C'est précisément sur ces aspects de l'évaluation des processus de mise en œuvre du programme que porte ce chapitre.

La partie de l'évaluation présentée ici comporte deux volets et poursuit quatre objectifs. Le tableau 17.1 présente une synthèse du plan d'évaluation concernant ces objectifs. Les objectifs liés à l'évaluation des effets du programme ont été retirés de ce tableau. Nous avons séparé, dans les volets I et II, les évaluations d'implantation et de processus afin de bien faire la distinction entre le caractère plus souvent descriptif associé à l'évaluation d'implantation et les éléments explicatifs et éclairants que peut apporter l'analyse des processus de mise en œuvre d'un programme. De fait, il n'existe pas de définition consensuelle de l'évaluation de processus ou d'implantation. Alors que certains auteurs utilisent les deux termes de façon interchangeable, d'autres limitent leur définition de l'évaluation de processus à la mesure des caractéristiques d'un programme et des clientèles qu'il rejoint, comparativement à ce qui était prévu par ses concepteurs (Rossi *et al.*, 2004). D'autres encore (Schreirer, 1994; Posavac et Carey, 2003) incluent dans leur définition de l'évaluation de processus les mesures d'implantation d'un programme dans une perspective de conformité (mesure des écarts entre ce qui a été mis en place comparativement à ce qui était prévu) et les mesures permettant d'expliquer ces écarts. Afin de bien distinguer ces deux aspects de l'évaluation liés à la mise en œuvre d'un programme, nous avons choisi, à l'instar de Patton

(1997) et Scriven (1991), de les présenter en deux volets intitulés respectivement «évaluation d'implantation» et «évaluation de processus». Le premier volet porte exclusivement sur les mesures de conformité d'un programme et repose sur des données quantitatives, alors que le second examine les mécanismes de mise en œuvre (référence des familles au programme et pratique des intervenants d'IRI-Accueil) et repose sur des données qualitatives.

L'évaluation poursuivait quatre objectifs :

- décrire le profil des familles desservies (volet I, implantation);
- décrire le degré d'implantation des huit principes directeurs (volet I, implantation);
- identifier les principaux éléments de la situation signalée qui mènent à la référence au programme (volet II, processus);
- mettre au jour des modèles implicites auxquels se réfèrent les intervenants dans leur pratique (volet II, processus).

Le premier volet de l'évaluation (implantation) visait deux objectifs, soit documenter les caractéristiques des familles et des enfants desservis par le programme IRI-Accueil afin d'en tracer le profil, et décrire les caractéristiques des services offerts. Ce second objectif sert à mesurer le degré d'application des principes directeurs qui définissent le cadre d'intervention du programme IRI-Accueil. L'ensemble des données recueillies pour répondre à ces deux premiers objectifs porte sur toutes les familles desservies entre le 29 novembre 2002 et le 13 juillet 2003.

Le deuxième volet de l'évaluation poursuivait deux objectifs : identifier les caractéristiques des situations qui mènent à un adressage au programme IRI-Accueil, et mettre au jour les modèles implicites auxquels se réfèrent les intervenants du programme. Pour chacun de ces deux objectifs, les données nécessaires à l'analyse proviennent d'entrevues individuels réalisés auprès de 15 intervenants des services Réception-traitement des signalements (RTS) et Urgences sociales (US) et de 11 intervenants du programme IRI-Accueil.

Nos choix méthodologiques comportent deux types de méthodes. Dans le volet I, nous avons fait le choix d'utiliser une **méthodologie quantitative** pour déterminer le degré d'implantation des huit principes directeurs et les caractéristiques des enfants et des familles. Les données nécessaires à ces analyses proviennent des systèmes d'information du CJM-IU et du DPJ.

TABLEAU 17.1
Synthèse des activités d'évaluation

Objectifs	Données recueillies	Variables	Sources des données	Participants
VOLET I : ÉVALUATION DE L'IMPLANTATION				
1. Décrire le degré d'implantation des principes directeurs	Caractéristiques des interventions	• Rapidité ; intensité ; souplesse horaire ; lieu où se déroule l'intervention ; concertation ; durée de l'intervention ; implication de toute la famille ; soutien concret.	• Formulaire maison remplis par les intervenants	Toutes les familles inscrites à IRI-Accueil entre le 29 novembre 2002 et le 13 juillet 2003 (n = 160)
2. Décrire le profil des familles desservies	Caractéristiques des enfants et des familles	• Signalements ; services ; placements six mois avant le signalement	• Banques de données du DPJ et du CJM-IU	
		• Enfants (âge, nombre d'enfants dans la famille, etc.)		
		• Parents (âge, structure familiale, etc.)		
		• Services reçus dans les six mois précédant le signalement		
VOLET II : ÉVALUATION DE PROCESSUS				
3. Identifier les caractéristiques de la situation	Processus d'adressage	• Éléments de la situation menant à la référence	Entretiens individuels semi-structurés	Intervenants RTS et Urgences sociales (n = 15)
4. Mettre au jour les modèles implicites auxquels se réfèrent les intervenants	Processus d'intervention	• Définition de la problématique	Entretiens individuels semi-structurés	Intervenants IRI-Accueil (n = 6 x 2 rencontres)
		• Activités mises en place		
		• Objectifs poursuivis		
		• Éléments favorables et obstacles		
		• Éléments de satisfaction ou d'insatisfaction		

Pour mener l'examen des *processus de référence* à IRI-Accueil (volet II), une **méthodologie qualitative** s'est avérée plus appropriée. Nous avons rencontré des intervenants des services RTS et US pour leur demander quels étaient les critères qu'ils utilisaient pour diriger ou non une famille vers le programme IRI-Accueil. Pour mener l'analyse des *processus d'intervention*, nous avons rencontré les intervenants d'IRI-Accueil à deux reprises pour leur demander de nous décrire les modalités et la nature de leurs interventions auprès de familles adressées par l'accueil du DPJ.

La description du processus d'évaluation

Nous avons mentionné plus haut que l'évaluation décrite dans ce chapitre découlait de l'initiative d'un groupe d'intervenants et de gestionnaires du CJM-IU. Ce groupe de personnes a été mis à contribution au moment de préparer la demande de subvention nécessaire à la réalisation de cette évaluation. Dès l'obtention des fonds, ce groupe s'est élargi afin d'inclure des membres du personnel de tous les paliers décisionnels du CJM-IU (intervenant du programme, gestionnaire de premier niveau, membre de la direction, conseiller professionnel) et des représentants des bailleurs de fonds. Le mandat de ce comité était de suivre le déroulement de l'évaluation afin d'aplanir les obstacles qui auraient pu surgir en cours de route et de transmettre à différents comités de gestion de l'établissement, dont le comité de direction, les résultats de l'évaluation au fur et à mesure où ils seraient disponibles. Les membres du comité ont également participé à l'interprétation des résultats. Bien qu'il ne s'agisse pas à proprement parler d'une évaluation participative, le rôle joué par ce comité s'apparente à celui des collaborateurs impliqués dans ce type d'évaluation, et sa participation active tout au long du processus explique largement l'utilisation des résultats de cette évaluation (Bilodeau *et al.*, 2006). Nous y reviendrons.

Un budget de 75 000 \$ a été obtenu pour la réalisation de l'évaluation. La subvention a été accordée conjointement par le Centre d'excellence pour la protection et le bien-être des enfants, la Fondation du Centre jeunesse de Montréal et le ministère de la Santé et des Services sociaux du Québec. L'essentiel de ce budget a servi à couvrir les salaires (étudiants diplômés et professionnels de recherche). Il est à noter que l'analyse des données qualitatives recueillies dans le cadre de cette étude a nécessité l'équivalent de quatre jours de travail par semaine pendant une période de neuf mois.

La collecte des données du volet I (quantitatif)

Les données nécessaires aux analyses de conformité proviennent de plusieurs sources que nous présentons ici.

- **Formulaires de chronologie des interventions.** Les intervenants du programme IRI-Accueil ont été mis à contribution afin de recueillir des données sur la nature et les modalités des interventions qu'ils effectuaient auprès des familles (nom de l'intervenant, date et heure de l'adressage, nom de l'enfant et numéro d'usager, date de l'intervention, heure de la première intervention, puis, pour chacune des interventions, l'indication de la période de la journée, la période de la semaine, la durée, l'offre de service concret, le lieu, les personnes présentes). Les formulaires de chronologie des interventions ont été remplis du 29 novembre au 17 décembre 2002 et du 10 mars au 31 juillet 2003 (dossiers fermés à cette date). Au total, 132 formulaires de chronologie des interventions ont été remplis.
- **Les systèmes d'information du CJM-IU et du DPJ.** Les informations sur les placements et les services offerts par le CJM-IU ainsi que les caractéristiques des jeunes et de leur famille (la localité où habite la famille, la date de naissance du jeune et de ses parents, la structure familiale et le sexe du jeune) proviennent du système d'information du centre jeunesse. Trois employés du centre ont compilé les données pour les 160 familles ayant bénéficié du programme IRI-Accueil pour la période du 29 novembre 2002 au 31 juillet 2003. Les informations sur les signalements, le nom du référent à l'accueil DPJ et le cadre légal de l'adressage proviennent du système d'information du DPJ. Une employée du centre jeunesse a compilé les données pour les 160 familles ayant bénéficié du programme IRI-Accueil pour la même période.
- **La base de données du programme IRI.** Afin de réduire au maximum le nombre de données manquantes, nous avons consulté la base de données du programme IRI-Accueil concernant les interventions qui ont eu lieu entre les deux périodes de collecte de données (soit du 18 décembre 2002 au 9 mars 2003) et pour lesquelles nous n'avons pas pu obtenir de formulaire de chronologie des interventions (soit 34). Pour ces cas, nous avons pu obtenir la date et l'heure de l'adressage à IRI-Accueil, le délai entre celui-ci et la première intervention, la durée totale des interventions en heures, le nombre total d'interventions, le

nombre d'interventions au téléphone, dans la famille et au bureau, ainsi que la date de fermeture du dossier.

- La base de données sur les familles. Elle contient les informations obtenues sur 160 familles ayant reçu les services du programme IRI-Accueil. Cette base comprend des données sur les caractéristiques de chaque famille, les critères d'adressage à IRI-Accueil ainsi que sur les services du centre jeunesse et les signalements dans les six mois précédant la référence à IRI-Accueil et les six mois suivant l'intervention. Des 165 cas ayant reçu les services IRI-Accueil de notre échantillon de départ, nous en avons retiré 5, puisque ces familles avaient fait l'objet de deux interventions IRI-Accueil pendant la période de collecte de données. Nous avons considéré la première intervention IRI-Accueil comme l'intervention cible, et la seconde comme un service du centre jeunesse offert dans les six mois suivant l'intervention.
- La base de données sur les interventions. Elle contient les informations portant sur les interventions individuelles dans le cadre de l'IRI-Accueil. Cette source inclut les données obtenues à partir du formulaire de chronologie des interventions rempli par les intervenants d'IRI-Accueil.

L'ensemble de ces données a ensuite été importé dans le système de traitement statistique SPSS (version 10.0), puis a été interrogé en fonction de nos hypothèses et questions d'évaluation (questions de départ et questions émergentes). Des tableaux d'inférences statistiques ont ainsi été produits, puis interprétés.

Les résultats quantitatifs de l'évaluation sont essentiellement descriptifs. La majorité des résultats se présentent sous la forme de moyennes, de fréquences et de proportions pour l'ensemble de l'échantillon et, dans certains cas, pour un sous-groupe de l'échantillon. Par exemple, pour le sous-groupe des interventions qui se sont produites la fin de semaine, quelle est la proportion des interventions effectuées le soir? Signalons le fait que le taux de données manquantes est parfois très élevé. Par exemple, il atteint 55 % pour les informations concernant la composition familiale.

La collecte des données du volet II (qualitatif)

Notre corpus qualitatif est composé de 11 entrevues réalisées avec les membres de l'équipe IRI-Accueil (une à deux rencontres par intervenant),

8 entrevues avec les intervenants RTS et 7 avec ceux de l'équipe US. Toutes ces rencontres ont eu lieu entre le 15 avril et le 28 août 2003¹.

Questions posées aux intervenants d'IRI-Accueil

- Pouvez-vous me parler de trois cas que vous avez suivis dans le cadre d'un adressage IRI-Accueil (décrire le contexte et la problématique)?
- Pourquoi avoir fait le choix de telle ou telle intervention?
- Quels sont les éléments qui ont nui à la situation et ceux qui ont eu un effet positif?
- Quelles ont été vos motifs de satisfaction et d'insatisfaction dans cette situation?

Questions posées aux intervenants de l'accueil DPJ

- Quels critères motivent l'adressage à IRI-Accueil?
- Quels critères excluent tout adressage à IRI-Accueil?
- Description du dernier cas adressé à IRI-Accueil et, pour ce cas précis, quels ont été les critères qui ont motivé l'adressage?
- Quelle différence y a-t-il entre adresser le jeune à IRI ou à un autre service?
- Qu'est-ce que cela donne de plus, pour la famille, pour le jeune?

Nous avons organisé et analysé le matériel à l'aide du logiciel de traitement de données qualitatives NVivo 2 (Q.S.R). Cet outil informatique nous a permis d'opérer un découpage des entrevues par thèmes (étape de la codification), d'élaborer une arborescence thématique et analytique (classification et hiérarchisation des catégories), puis d'effectuer des croisements, des associations et des recoupements afin de tester nos hypothèses et de comparer certains segments de discours (Giordano, 2003).

Bien sûr, nous avons organisé le matériel autour des grands thèmes que nous souhaitions explorer davantage, mais nous avons été particulièrement attentifs aux thèmes émergents (par exemple, engagement des intervenants auprès des familles, thème de la loyauté, conception d'une intervention IRI-Accueil réussie). Autrement dit, à partir de questions ouvertes, nous nous sommes efforcés d'identifier les principaux « objets » qui revenaient dans les discours (thèmes, concepts, mots...) et de comprendre au besoin les conditions de leur présence (par exemple, à proximité du vocable

1. Les six intervenants de l'équipe IRI-Accueil ont été rencontrés. Huit intervenants RTS sur neuf et sept intervenants US sur neuf ont été rencontrés.

« crise », on trouve généralement des mots tels que « colère », « intervenir vite », etc.), pour ensuite émettre des hypothèses (par exemple, toutes les personnes interrogées n'accordent pas le même sens au concept de crise), puis les vérifier, le cas échéant (recherche lexicale autour du mot « crise », examen de la « mise en scène » du concept de crise...).

Pour vérifier la validité de nos hypothèses et de nos interprétations, nous avons soumis le matériel aux instances concernées pour obtenir leur avis. Ainsi, dans le cadre d'une réunion prévue à cet effet, nous avons présenté les résultats de l'évaluation aux membres des équipes RTS, Urgences sociales et IRI-Accueil en vue de recueillir leurs commentaires. Un peu plus tard, nous avons remis une copie du rapport à quelques volontaires (IRI-Accueil, RTS et Urgences sociales) ainsi qu'aux membres du comité de suivi pour obtenir leurs commentaires.

Le matériel d'analyse est, nous l'avons vu, constitué d'entrevues et de données statistiques, lesquelles ont été recueillies entre novembre 2002 et juillet 2003. Cela dit, au fil des mois, d'autres informations nous sont parvenues au cours de consultations avec les membres du comité de suivi. Celles-ci nous ont permis de mieux comprendre le contexte et certains enjeux des services de l'accueil DPJ et de IRI-Accueil, mais aussi de mettre en contexte nos descriptions et nos analyses, et de les inscrire dans le temps. Pour respecter les mesures concernant l'anonymat des personnes (les informateurs), lorsque ces informations étaient présentées dans le rapport d'évaluation, nous avons précisé qu'elles avaient été recueillies dans le cadre de consultations complémentaires.

Les mesures de confidentialité

Plusieurs mesures ont été employées pour assurer la confidentialité des données recueillies. Les répondants ont signé un formulaire de confidentialité et les extraits présentés dans le rapport ont été rendus anonymes par le brouillage des indicateurs de sexe et des mots ou phrases pouvant conduire à l'identification des personnes. Ajoutons que tous les noms et prénoms qui figurent dans les extraits d'entrevue présentés dans le rapport sont fictifs. Enfin, toutes les données recueillies sont entreposées selon les normes en vigueur à l'Université de Montréal.

Afin qu'il soit fidèle à l'esprit d'évaluation participative qui a guidé cette démarche, le rapport a été rédigé de manière à ce que les personnes

les plus directement touchées par l'étude, soit les intervenants et gestionnaires du Centre jeunesse de Montréal – Institut universitaire, puissent s'approprier facilement les résultats les plus utiles pour elles. Pour cela, nous n'avons présenté dans le rapport que les résultats les plus marquants tirés de nos analyses. Des résultats plus détaillés ainsi que toutes les informations traditionnellement exigées dans les publications scientifiques (méthodologies, références, questionnaires, stratégies statistiques, etc.) sont disponibles dans des appendices regroupés dans un autre document. Ils sont donc accessibles aux personnes qui désirent connaître ces informations complémentaires. Le rapport se termine par une discussion approfondie qui, au lieu d'aboutir à des recommandations souvent considérées comme prescriptives par les personnes responsables dans les établissements de services, suggère plutôt des pistes d'action. Nous y reviendrons dans la dernière section de ce chapitre.

Les résultats

Les caractéristiques des interventions IRI-Accueil

À la lumière de nos analyses, les résultats montrent que l'intervention IRI-Accueil respecte six des huit principes directeurs du modèle auquel elle se réfère. Par contre, deux caractéristiques du modèle s'écartent de ces principes directeurs. D'abord, pour l'ensemble de nos données (tant quantitatives que qualitatives), il n'est pratiquement pas fait mention d'activités de soutien concret auprès des jeunes ou des familles, plus des deux tiers des interventions se prolongent au-delà de la durée prévue.

Les modèles d'intervention implicites des professionnels du programme IRI-Accueil

Concernant les durées moyennes des interventions, très souvent supérieures à ce qu'elles devaient être, il se dégage de nos analyses que ce phénomène était intimement lié à la question du transfert des dossiers aux partenaires; certains intervenants semblent utiliser ces ressources plus rapidement ou plus fréquemment que d'autres. Par ailleurs, il ne semble pas aisé, pour les intervenants, de rompre le lien avec les familles et d'envisager l'aide à leur apporter sur une base de très court terme.

L'évaluation des processus de référence : l'intervention des équipes RTS et US

Les cas adressés au programme IRI-Accueil qui nous ont été décrits apparaissent généralement très pertinents. Compte tenu de la complexité des situations traitées, il ne fut pas aisé de dégager des critères permettant de prendre avec certitude la décision d'adresser ou pas un cas à IRI-Accueil. En ce qui concerne le *critère de crise*, pour central qu'il soit, il n'apparaît pas réellement opérationnel pour aider à la prise de décision, car il s'agit d'un concept fortement sujet à interprétation. Nous constatons également que le fait de parler de *troubles de comportement* sans en spécifier la nature ni l'intensité n'est pas suffisant pour évaluer une situation ou prendre une décision, et il semble particulièrement difficile d'estimer le seuil de gravité des comportements pour lesquels IRI-Accueil peut être ou non efficace.

Cela dit, les membres des équipes RTS et US, en raison de leur grande expérience en évaluation de situations, sont en mesure d'indiquer toute une série de critères qui les aident considérablement à faire le choix d'adresser un jeune ou pas à IRI-Accueil. En général, le programme IRI-Accueil n'est pas mobilisé quand les troubles de comportement du jeune sont graves ou sérieux, quand le jeune est déjà suivi par un autre partenaire ou quand il est connu des services du DPJ. En revanche, le recours à ce programme est généralement envisagé pour éviter le placement ou la rétention de signalements, quand on juge qu'il est nécessaire d'intervenir rapidement parce que les membres de la famille sont à bout, pour éviter que la situation ne se détériore, quand la famille n'a pas la capacité ou la volonté d'aller chercher de l'aide, ou encore parce qu'une intervention au cœur de la crise augmente les chances d'aider réellement la famille. Tous les intervenants de l'accueil DPJ n'ont pas la même perception de l'efficacité des CLSC² et n'adressent donc pas de cas à cette ressource externe selon les mêmes conditions. Pour des situations similaires, certains préféreront mobiliser l'équipe IRI-Accueil plutôt que d'orienter les familles vers les CLSC, d'autres non. Certains semblent adresser au programme IRI-Accueil des cas « coriaces » alors que d'autres, des cas plus « légers ». La collaboration des familles est aussi un critère très souvent mis de l'avant pour motiver et justifier le recours au programme IRI-Accueil.

2. CLSC : centre local de services communautaires.

Discussion

Les limites de l'évaluation

Il convient de mentionner que le nombre élevé de données manquantes concernant certaines variables a certainement limité la portée et l'interprétation des résultats de cette évaluation. Ainsi, dans les bases de données du DPJ et du centre jeunesse, il ne nous a pas toujours été possible de trouver l'information nécessaire à propos de variables comme la structure familiale, l'âge des parents, l'origine ethnique de la famille et le nombre de personnes vivant au domicile familial. De minces écarts ont parfois été observés pour des données communes mais traitées par différents services (par exemple, l'heure de la référence d'un cas par l'accueil DPJ qui figure à la fois dans les formulaires de chronologie et dans les bases de données du DPJ). De plus, des données n'ont pas toujours été répertoriées dans les questionnaires remplis par les intervenants IRI-Accueil, et des indices nous portent à croire que certaines l'ont été de façon erronée. À titre d'exemple, nous avons repéré trois formulaires de chronologie des interventions ayant été remplis deux fois pour une même famille, mais avec des entrées différentes. Par ailleurs, les formulaires de chronologie des activités nous parvenaient souvent plusieurs semaines après la fin des interventions. Pour que les données aient été vraiment fiables, il aurait fallu que les intervenants les recueillent de façon systématique. Nous croyons qu'un suivi hebdomadaire devrait être effectué, lors de telles collectes de données, afin d'éviter que certains éléments échappent à la mémoire.

Il faut mentionner le fait que nous n'avons pas réussi à documenter les interventions menées par d'autres organismes ou établissements que le CJM-IU pendant et à la suite de l'intervention de l'équipe IRI-Accueil. Cela aurait permis d'aller beaucoup plus loin dans l'analyse de l'efficacité de ce programme, notamment en ce qui concerne la qualité de la collaboration entre chacun des partenaires et la complémentarité de leurs interventions. Cet examen, envisagé lors de la planification de l'évaluation, avait été écarté non pas parce qu'il manquait d'intérêt, mais simplement parce qu'il aurait exigé des ressources financières considérables dont l'organisme ne disposait pas.

Une approche axée sur la mise à profit des résultats

Au total, une série de 14 pistes d'action ont été suggérées dans le rapport final, toutes appuyées sur l'analyse de nos données et soutenues par les membres du comité de suivi. Nous présenterons ici deux exemples de ces pistes d'action.

La première découle du fait qu'au moment de recueillir les données, il existait peu d'écrits concernant les modalités et le processus de référence de cas à IRI-Accueil. Au cours des huit mois qu'a duré la collecte des données quantitatives, certains intervenants ont adressé 15 cas au programme IRI-Accueil, d'autres, aucun ou un seul. Nous avons donc examiné les taux d'adressage à IRI-Accueil (en tenant compte du nombre d'heures de présence des intervenants et en excluant les heures de nuit) pour constater que certaines personnes en poste seulement trois ou quatre heures par semaine, en moyenne, avaient adressé plusieurs cas, alors que d'autres, en poste à mi-temps, n'en avaient adressé aucun. Or, à la suite de la présentation des résultats préliminaires issus de nos analyses, un « Protocole d'intervention au programme I.R.I » a été rédigé conjointement par le chef de service et le coordonnateur professionnel du programme. Ce protocole, à l'usage des intervenants RTS et Urgences sociales, est diffusé depuis mars 2004. Il éclaire certainement le mandat d'IRI-Accueil.

La seconde porte sur l'application des principes directeurs du programme. Nos résultats montrent que la durée de prise en charge était en moyenne beaucoup plus longue que celle qui était initialement prévue dans les principes du programme. L'analyse des discours a montré que des éducatrices ou éducateurs semblaient encore fortement attachés à leur rôle d'agent de changement ou encore à ce sentiment subjectif d'avoir aidé les familles. Il n'est pas malaisé de comprendre cet état de fait lorsque l'on considère l'histoire professionnelle de la plupart de ces intervenants, habitués qu'ils étaient à mener leurs interventions jusqu'à leur terme. Il y a aussi la croyance, partagée par plusieurs, qui prête à l'intervention longue des vertus qu'elle ne semble pas avoir dans le contexte du programme IRI-Accueil. Les résultats quantitatifs ont en effet montré qu'une durée prolongée du suivi n'empêche pas significativement les jeunes d'être de nouveau signalés, de recevoir d'autres services du centre jeunesse ou d'être placés à la suite de l'intervention. Nous avons donc proposé de

mener des discussions au sein de l'équipe IRI-Accueil et avec ses responsables afin de trouver un consensus sur les critères de réussite du programme, et cela pour aider les éducatrices et éducateurs à valoriser la portée de leurs interventions et à mieux saisir les effets escomptés du programme. Parallèlement, nous suggérons de poursuivre le monitoring et l'analyse périodique de la durée des interventions.

Au moment de déposer notre rapport, ces actions, ainsi que plusieurs autres, étaient déjà amorcées depuis plusieurs mois. Afin de donner suite aux conclusions de cette évaluation, nous avons suggéré que le comité de suivi, qui a assuré l'adéquation entre les modalités d'évaluation et les besoins du CJM-IU en matière d'information sur le programme, soit officiellement mandaté pour instaurer des mesures correspondant aux pistes d'action formulées dans ce rapport et en assurer le suivi.

Nous avons parlé, dans l'introduction, de la création du comité de suivi et de son rôle central dans la mise en œuvre de l'évaluation et dans sa réalisation. Cette stratégie s'inscrivait clairement dans une approche orientée sur l'utilisation des résultats de l'évaluation, et elle a donné lieu à plusieurs retombées concrètes. Par exemple, tout au long de la démarche, les résultats ont été présentés au comité de suivi, dont certains membres, siégeant également à divers comités (direction, groupe de développement clinique, etc.), étaient en bonne position pour suggérer d'inviter les chercheurs à présenter les résultats aux moments jugés opportuns. Ces présentations par vagues successives ont permis de procéder à des ajustements du programme au fur et à mesure que les résultats émergeaient. Cela a fait en sorte que notre rapport final, qui faisait la lumière sur certains aspects négatifs de la mise en œuvre du programme, notamment en ce qui concernait la durée moyenne des interventions (nettement plus longue que prévu) et la disparité dans l'interprétation des critères de référence au programme, a été accueilli sans fracas : la plupart des mesures correctives avaient déjà été envisagées par les parties prenantes. Cette façon de faire a certainement aussi largement contribué à renforcer au sein de cet établissement de services la conviction que l'évaluation de programmes peut permettre d'améliorer les pratiques et la prise de décision concernant les programmes.

18

Une évaluation des processus d'un programme de référence-évacuation au Mali

*Pierre Fournier, Caroline Tourigny
et Francine LaBossière*

Ce chapitre traite de l'évaluation d'un programme de référence-évacuation des urgences obstétricales dans un pays à faibles ressources, le Mali. Cette évaluation, qui est toujours en cours, consiste en une analyse de l'implantation et des effets du programme. Elle répond aux besoins d'information des acteurs locaux et régionaux qui ont participé activement à sa définition et à sa mise sur pied, et ce, dans une région précise du Mali. Nous décrirons tout d'abord le programme ainsi que le contexte dans lequel il s'inscrit. Nous présenterons ensuite succinctement les principales approches utilisées dans l'analyse de l'implantation et le protocole d'évaluation retenu. Enfin, nous décrirons en détail le processus d'évaluation et présenterons quelques-uns des premiers résultats et les leçons apprises jusqu'à ce jour.

Le contexte et le programme

Le Mali est situé en Afrique de l'Ouest, et sa population est de 14 517 176 habitants (Institut national de la statistique, 2009), dont plus de 70 % vivent en zone rurale. L'espérance de vie à la naissance est de 61,6 ans. La majorité de la population vit dans la pauvreté, et 35 % dans l'extrême pauvreté. Le

milieu rural enregistre généralement une incidence et une profondeur de la pauvreté plus forte que le milieu urbain selon l'Enquête démographique et de santé du Mali 2006 (Cellule de planification et de statistiques du ministère de la Santé *et al.*, 2007).

Le système de santé comprend quatre niveaux : les centres de santé communautaires (CSCoM), les centres de santé de référence de district (CSRéf) et les hôpitaux régionaux et nationaux. Les CSCoM constituent le premier niveau de contact avec le système de santé. C'est la population qui les gère par l'entremise de l'Association de santé communautaire (ASACO), et ils offrent des soins de base tels que la consultation prénatale, la vaccination, des consultations curatives et des accouchements normaux. Ils s'adressent à une population de 5 000 à 15 000 habitants. Pour assurer la continuité des soins, chaque CSCoM est en relation avec un CSRéf vers lequel il dirige les cas qui dépassent sa compétence.

La région de Kayes est une des huit régions administratives et sanitaires du Mali. Elle compte 1 996 812 habitants (Institut national de la statistique, 2009) et est divisée en sept districts groupant 1 565 villages. La densité de la population se situe entre 9 et 21 habitants au kilomètre carré. On trouve un hôpital régional à Kayes, la capitale de la région. Chaque district compte un CSRéf, et l'on y dénombre 185 CSCoM et 62 autres centres de santé de premier niveau (postes de santé, dispensaires) (DRS et DRDSES-Kayes, 2011). Ces derniers centres ne font pas partie du système étatique et sont gérés par des collectivités, des organisations non gouvernementales ou des organisations religieuses.

Malgré le développement soutenu du secteur sanitaire au cours des dernières années, la couverture sanitaire demeure faible : en 2010, 49 % de la population habitait à plus de 5 km d'un centre de santé, et 26 % à plus de 15 km (DRS et DRDSES-Kayes, 2011).

En 2010, la proportion des accouchements assistés dans la région atteignait 56 % (DRS et DRDSES-Kayes, 2011).

Le taux de mortalité maternelle¹, au Mali, est un des plus élevés au monde. Le ratio de mortalité maternelle serait de 464 décès pour 100 000 naissances vivantes (Cellule de planification et de statistiques du ministère

1. Un décès maternel a été défini comme le décès d'une femme survenant au cours de la grossesse ou dans les 42 jours suivant l'accouchement, pour une cause obstétricale directe ou indirecte (Classification statistique internationale des maladies, neuvième révision, OMS).

de la Santé *et al.*, 2007). Une estimation de ce ratio corrigeant les biais de mesure le situerait à 830 sur 100 000 pour l'année 2008 (WHO, 2010). La réduction du taux de mortalité maternelle est un des objectifs du Millénaire des Nations Unies pour le développement; on vise en effet à la réduire de 75 % entre 1990 et 2015 (Projet Objectifs du Millénaire des Nations Unies, 2005), ce qui représente un défi de taille pour le Mali, qui a plutôt fixé son objectif de réduction à 50 % (Ministère de l'Économie et des Finances, Mali, 2002).

Causes, conditions et stratégies de réduction de la mortalité maternelle

Les causes et les conditions de la mortalité et de la morbidité maternelles sont bien connues, et les stratégies permettant de la réduire ont fait l'objet d'un recensement exhaustif (Ronsmans et Graham, 2006). On estime que la majorité des décès maternels surviennent entre le début du dernier trimestre et la semaine suivant la fin de la grossesse (Campbell et Graham, 1990; Li *et al.*, 1996). La plupart des décès maternels sont des conséquences directes de la grossesse et de l'accouchement. Plus précisément, les femmes meurent principalement d'hémorragies sévères, de désordres hypertensifs et d'infections (Khan *et al.*, 2006), mais les avortements, surtout ceux qui sont provoqués, demeurent une cause majeure de décès (Thonneau *et al.*, 2002).

Une des stratégies de réduction de la mortalité maternelle qui a fait la preuve de son efficacité vise à garantir aux femmes enceintes l'accès à des soins obstétricaux de base (SOU) – administration parentérale d'oxytociques, d'antibiotiques et d'anticonvulsivants, accouchements assistés, extraction manuelle du placenta – et complets, ou les SOU de base avec césarienne et transfusion (Bernis, 2003). Une revue systématique de la littérature (Paxton *et al.*, 2005) montre qu'un meilleur accès aux SOU est associé à des taux de mortalité maternelle moins élevés, mais que la qualité des soins qu'on y dispense est tout aussi importante, l'accès facile à des SOU de mauvaise qualité n'entraînant pas la même réduction de la mortalité maternelle. Les auteurs ont démontré que les données sont concluantes et que les SOU doivent faire partie de tout programme qui a pour but de réduire la mortalité maternelle.

L'accès aux SOU demeure cependant problématique dans des contextes à faibles ressources. Dans un pays en développement, les femmes qui ont une complication obstétricale doivent souvent surmonter plusieurs obs-

tacles pour pouvoir accéder aux SOU. Le modèle des trois retards, ou des trois délais, élaboré par Thaddeus et Maine (1994), divise le temps qui s'écoule entre l'apparition d'une complication et l'obtention des soins obstétricaux d'urgence appropriés. Les trois retards décrits par ces auteurs sont :

- le temps écoulé avant de décider d'avoir recours aux SOU ;
- le temps nécessaire pour se rendre à un centre de SOU ;
- le temps écoulé entre l'arrivée au centre de SOU et la dispensation des soins appropriés.

Parmi les facteurs qui peuvent influencer sur le premier retard, correspondant à la décision d'avoir recours aux SOU, on trouve la non-reconnaissance des signes de danger, la gravité de la complication, l'éloignement du centre de santé, l'appréciation de la qualité des soins offerts et le refus de la femme ou de la famille d'avoir recours aux SOU pour des raisons économiques ou socioculturelles. Le deuxième retard est souvent lié à la situation géographique des centres de santé, à leur éloignement par rapport au domicile, aux conditions des routes et à la disponibilité et aux coûts du transport. Le troisième retard peut être dû au système d'adressage des SOU de base vers les SOU complets, système qui peut être inadéquat ou manquer de personnel compétent ou de matériel médical (Maine *et al.*, 1997).

Selon différentes études (répertoriées par Ronsmans et Graham, 2007), les femmes meurent majoritairement en milieu hospitalier, ce qui inclut des cas ayant été adressés pour complications obstétricales et des femmes venues accoucher normalement mais qui ont connu des complications au cours du travail. La qualité des soins demeure médiocre, et plusieurs études ont fait état de pratiques cliniques contestables et de délais dans le diagnostic et le traitement des complications (Harrison, 1997 ; cité par Ronsmans et Graham, 2006). Selon une étude multisite effectuée au Sénégal, 79 % des décès survenus dans les centres hospitaliers étaient évitables et étaient attribuables à la mauvaise qualité des soins (Dumont *et al.*, 2006).

Description du programme

Pour réduire le taux de mortalité maternelle, le gouvernement du Mali a mis en place un système de référence-évacuation (SRE) pour toute urgence obstétricale, qu'on définit comme un état grave lié à la grossesse ou à l'accouchement et requérant une intervention médicale d'urgence (Maine *et al.*, 1997). Ce système de référence vise à améliorer la qualité des soins obstétricaux d'urgence ainsi que l'accès des femmes à ces soins, de façon à diminuer la létalité due aux complications obstétricales. *Grosso modo*, il comporte trois composantes principales: un système de financement (des caisses de solidarité); un système d'alerte et d'évacuation sanitaire (des ambulances et des radios); la mise à jour des soins obstétricaux d'urgence dans les CSCom et les CSRéf.

- *Les caisses de solidarité.* Présentes dans chaque district, elles ont pour but de réduire les coûts assumés par les femmes lors d'une complication. Des partenaires locaux (ASACO, Mairies, CSRéf) font des contributions financières à la caisse sur une base régulière; la contribution de chaque partenaire est établie annuellement par consensus. L'argent de la caisse est utilisé en cas d'évacuation, pour payer une partie des coûts directs et indirects (par exemple, le carburant de l'ambulance, l'allocation journalière du chauffeur, l'amortissement de l'ambulance, les soins fournis). Avant l'établissement de ces caisses, les patientes devaient payer la totalité des coûts, ce qui constituait une barrière financière importante et pouvait retarder de façon considérable l'obtention de soins appropriés.
- *Le système d'alerte et d'évacuation.* Chaque CSCom possède une radio rurale permettant de contacter le CSRéf. On utilise aussi de plus en plus les téléphones cellulaires pour communiquer avec le CSRéf. Chaque CSRéf dispose d'au moins une ambulance pouvant assurer l'évacuation des urgences obstétricales du CSCom vers le CSRéf, une fois l'appel lancé par le CSCom.
- *Les soins obstétricaux d'urgence (SOU) à deux niveaux.* Des activités de formation et des achats d'équipements ont permis d'améliorer la disponibilité et la qualité des SOU. Ceux-ci comprennent deux niveaux. Les SOU₁, dans les CSCom, prennent en charge les accouchements normaux, déterminent les complications et dirigent les cas graves vers le CSRéf. La majorité des CSCom sont dirigés par un infirmier, alors

qu'une matrone s'occupe généralement des accouchements. Quelques CSCom, dirigés par des médecins ayant reçu une formation spécifique en SOU de base, peuvent accomplir des actes plus complexes tels que l'installation de perfusions d'oxytocines et l'application de ventouses ou de forceps. Les SOU₂, dans les CSRéf, incluent les SOU₁ et effectuent en plus les accouchements compliqués et de la prestation des SOU complets. Les CSRéf sont dotés d'un bloc opératoire, et la plupart d'entre eux peuvent faire des transfusions sanguines.

L'intervention est relativement normalisée du point de vue technique et bien décrite dans un cadre conceptuel national, mais ce cadre doit ensuite être adopté par chaque district. Pour que le système de référence-évacuation soit considéré comme effectif dans un district donné, la communauté et les partenaires du district doivent s'entendre sur les responsabilités de chacun. Le SRE est actuellement effectif dans les sept districts de la région de Kayes, mais la date de mise en œuvre du système varie d'un district à l'autre, elle se situe entre juillet 2001 et novembre 2005.

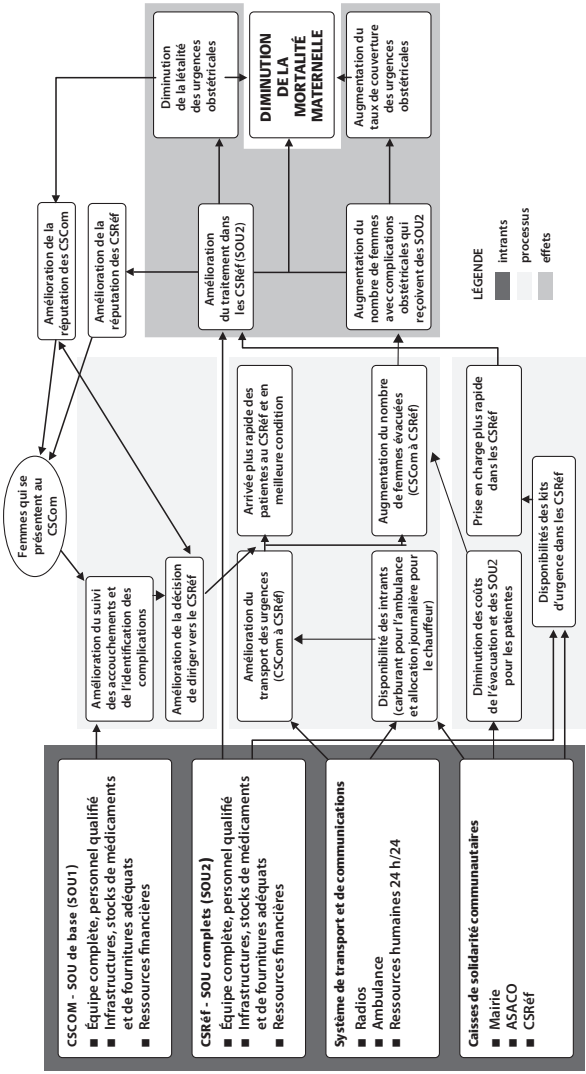
Le modèle logique de l'intervention

Le modèle logique du programme est illustré à la figure 18.1. Les différents intrants y sont représentés (CSCom, CSRéf, systèmes de transport et de communication, et caisses de solidarité), ainsi que leurs liens avec les différents processus menant à la production des effets attendus du SRE.

Description des acteurs et des enjeux de l'évaluation

Différents acteurs participent directement ou indirectement au programme du système de référence-évacuation, et cela à plusieurs niveaux : local, régional, national, supranational et international. Les enjeux de l'évaluation varient en fonction du niveau où travaille chaque acteur et de son rôle dans le SRE, comme le montre le tableau 18.1.

FIGURE 18.1
Modèle logique du système de référence-évacuation



TABEAU 18.1**Acteurs et enjeux dans l'évaluation du système de référence-évaluation**

Acteurs	Rôles	Enjeux de l'évaluation
<i>Niveaux local et régional</i>		
Direction régionale de la santé (DRS)	Mise en œuvre des politiques nationales en santé (pour la DRS) et sociales (pour la DRDSES).	Il s'agit de la première région sanitaire à mettre en place un système de suivi du SRE aussi complet.
Direction régionale du développement social et de l'économie solidaire (DRDSES)		Des données probantes sont recueillies sur les effets du SRE à l'échelle régionale et sur les facteurs influant sur son implantation.
Équipe sociosanitaire de district (médecins du CSRéf et personnel du développement social et de l'économie solidaire)	Mise en œuvre des politiques sociales et en santé, à l'échelle des districts.	Ces données permettent de révéler les résultats tangibles du programme au bailleur de fonds (l'Agence canadienne de développement international) ainsi qu'au ministère malien de la Santé. Ces résultats permettent d'améliorer le SRE et de maximiser son potentiel, et donc de convaincre les collectivités locales de la pertinence de soutenir le SRE et de cotiser selon le montant prévu aux caisses de solidarité.
Communautés et collectivités locales	Promotion et mise en opération des systèmes locaux de santé. Contribution financière des caisses de solidarité. Fourniture par le SRE de services à la population.	Des résultats concluants sur la diminution de la létalité maternelle confirmeront la pertinence de l'intervention.
<i>Niveau national</i>		
Direction nationale de la santé (DNS)	Appui et suivi de la mise en œuvre des politiques nationales de santé.	Collecte et production de données probantes sur les effets du SRE et sur les facteurs qui influent sur son implantation, ce qui peut s'avérer utile pour motiver les autres régions sanitaires à créer ou à étendre leur propre SRE, tout en informant sur les conditions optimales de création.
Cellule planification et statistique (ministère de la Santé)	Responsable de la conception des politiques nationales en santé.	
<i>Niveau supranational</i>		
Organisation mondiale de la Santé (OMS)	Autorités mondiales concernant les objectifs du Millénaire pour le développement.	Production de preuves sur l'efficacité des interventions visant à réduire la mortalité maternelle afin de diffuser les meilleures pratiques.
Fonds des Nations Unies pour la population (UNFPA)		

Acteurs	Rôles	Enjeux de l'évaluation
<i>Niveau international</i>		
Université de Montréal	Recherche en santé mondiale ainsi qu'interventions en santé. Élaboration d'un programme de recherche et de formation.	Évaluation globale mais rigoureuse d'une intervention en santé publique qui porte sur un problème majeur en santé mondiale. Terrain de formation de qualité pour des étudiants africains et canadiens. Diffusion dans la communauté scientifique des résultats sur les effets positifs du SRE. Établissement d'une relation solide et à long terme avec la DRS et la DRDSES-Kayes, fondée sur l'échange et la confiance, et sur laquelle s'appuiera un programme de recherche plus large.
Agence canadienne de développement international (ACDI)	Appui financier et apport d'expertise aux DRS et DRDSES-Kayes (2003-2010).	Production de données probantes sur les effets d'un programme appuyé par l'ACDI.
Centre de recherches pour le développement international (CRDI, Canada)	Attribution de fonds pour la recherche.	Appui aux recherches liées à l'intervention, facilitant la prise de décision et associant chercheurs et décideurs. Valorisation de la recherche effectuée dans un contexte de transfert des connaissances et de renforcement des capacités.

Approches dans l'analyse d'implantation et protocole d'évaluation retenu

Il s'agit ici de décrire l'analyse de la production des effets d'une intervention complexe dont la mise en œuvre est largement influencée par divers facteurs. Ce type d'évaluation est connu sous le nom d'analyse d'implantation ou encore d'évaluation de processus. Bien que tous les auteurs s'entendent sur le fait que l'analyse d'implantation s'oppose au modèle de la « boîte noire » et qu'elle a pour but de comprendre les mécanismes internes par lesquels les interventions produisent leurs effets en tenant compte du contexte, il n'existe pas de modèle universel d'analyse d'implantation. On peut donc distinguer différentes approches, mais un certain nombre d'éléments sont communs à toutes. Décrivons brièvement cinq de ces approches.

L'approche suivie par Champagne et Denis (1990) distingue deux genres de processus d'évaluation ; l'évaluation normative et la recherche évalua-

tive. L'analyse d'implantation se rattachant à celle-ci comporte les trois composantes suivantes :

- l'analyse des déterminants contextuels du degré de mise en œuvre de l'intervention. Elle vise à déterminer les facteurs qui expliquent les différences entre l'intervention planifiée et l'intervention mise en œuvre. Elle sert aussi à reconnaître les milieux où une implantation intégrale de l'intervention est possible ;
- l'analyse de l'influence des variations dans le processus d'implantation sur les effets observés ; elle a pour but de déterminer l'influence de chaque degré de mise en œuvre (variables indépendantes) sur les effets produits par l'intervention (variables dépendantes). Elle permet de discerner les composantes essentielles et secondaires des interventions pour l'obtention des effets attendus ;
- l'analyse de l'influence de l'interaction entre le contexte d'implantation et l'intervention elle-même sur les effets observés. Elle permet de comprendre l'interaction (synergie ou antagonisme) entre différents facteurs en jeu dans la production des effets. Cette analyse est utile pour documenter la dynamique interne d'une intervention.

Patton (1997) distingue cinq dimensions dans l'analyse d'implantation :

- l'évaluation de l'effort, qui a pour but d'apprécier la quantité et la qualité des ressources d'un programme et de déterminer dans quelle mesure le programme est actif ;
- le monitoring des programmes, qui a pour objet d'assurer la mise en place d'un système de collecte de données de routine ;
- l'analyse des processus, qui a pour but de déceler les forces et faiblesses d'un programme au moyen d'une analyse de la dynamique interne et de ses composantes ;
- l'analyse des composantes, qui a pour rôle d'évaluer les diverses parties d'un programme. En d'autres mots, un programme pouvant être la somme de plusieurs efforts opérationnels, chacun de ceux-ci peut faire l'objet d'une analyse d'implantation. Les liens entre une ou plusieurs de ces composantes peuvent aussi faire l'objet d'une évaluation ;
- la spécification du traitement, qui a pour but de définir et de mesurer chaque élément du programme ainsi que l'effet qui en est attendu. Elle a aussi pour rôle d'établir les liens de causalité et de valider la théorie de l'intervention.

Pour Rossi, Lipsey et Freeman (2004), l'analyse d'implantation est non pas une simple procédure ou une série de procédures, mais plutôt une étude visant à trouver des réponses à des questions qui concernent à la fois l'intervention en tant que telle (ressources, opérations, composantes et activités), ses fonctions et son déroulement. Pour ces auteurs, cette forme d'évaluation constitue un complément indispensable à l'évaluation d'impact.

Pawson et Tilley (1997, 2005) prônent pour leur part une évaluation dite réaliste (*realistic*) mettant l'accent sur quatre concepts clés interdépendants (voir chapitre 14) qui permettent d'expliquer et de comprendre les programmes :

- les mécanismes de production des effets se rapportent à la façon dont une des composantes, un ensemble de celles-ci ou encore une succession d'étapes conduisent à des changements. Ces mécanismes constituent la logique d'une intervention ;
- le contexte décrit les conditions dans lesquelles le programme est mis en œuvre et qui influent sur les mécanismes de production des effets. Le contexte inclut, bien sûr, le lieu d'intervention et aussi, entre autres, les relations interpersonnelles et sociales, la biologie, les technologies, les conditions économiques ;
- les effets englobent à la fois les conséquences attendues ou non attendues d'un programme et résultant de la mise en action de différents mécanismes dans différents contextes ;
- le modèle contexte-mécanismes-effets consiste à observer comment les variations dans les mécanismes et les contextes peuvent permettre de prédire et d'expliquer les variations dans les résultats d'une intervention. Les conclusions d'une évaluation réaliste touchent toujours la configuration des différentes composantes nécessaires au maintien d'un programme.

Chen (2005) prône pour sa part une évaluation des processus basée sur la théorie de l'intervention (*theory-driven process evaluation*) et une évaluation intégrée de processus et des effets (*integrative process/outcome/evaluation*). Ces deux types d'analyse se basent sur la théorie du programme, soit les hypothèses causales qui sous-tendent l'implantation du programme et la production des effets. Le premier évalue le degré d'accord entre la théorie sur laquelle s'appuie le programme et son implantation

réelle, tandis que le second analyse les suppositions sous-jacentes des effets qui seront produits.

L'analyse d'implantation du système de référence-évacuation du Mali s'inscrit plutôt dans la perspective de Chen, qui correspond à l'évaluation basée sur la théorie. Nous avons privilégié cette approche qui exige de rendre explicite la théorie du programme en raison de la complexité particulière du SRE et du nombre de niveaux où agissent les acteurs. Il était nécessaire, selon nous, de commencer par la construction du modèle logique de l'intervention, résumé à la figure 18.1. Par ailleurs, les acteurs ne voient pas tous nécessairement le SRE de la même façon ; ainsi, le fait de rendre la théorie de l'intervention explicite par la construction d'un modèle logique nous permet de nous entendre sur l'objet de l'évaluation. Cette clarification initiale vise à faciliter le déroulement ultérieur de l'évaluation et les échanges entre les acteurs durant tout le processus.

L'analyse d'implantation basée sur la théorie de l'intervention

Il est clair que la mortalité maternelle est un problème de santé majeur dans les pays en développement. C'est pour cette raison que les gouvernements d'un grand nombre de ces pays ont élaboré et mis en œuvre des politiques et des stratégies visant l'amélioration de la santé maternelle. Des solutions simples, efficaces et relativement peu coûteuses existent pour réduire le taux de mortalité maternelle, notamment l'amélioration de la qualité des soins obstétricaux d'urgence et de l'accès à ces SOU. Par contre, dans plusieurs régions où de telles interventions ont été mises en place, on n'obtient pas les résultats escomptés. Les interventions basées sur les SOU, qui ont pourtant remarquablement réussi à faire chuter les taux de mortalité maternelle dans les pays industrialisés, ne connaissent pas le même succès dans les pays en développement, car les mécanismes de production des effets sont toujours méconnus dans ces derniers.

En rendant explicites les hypothèses causales sous-jacentes à l'intervention et en construisant son modèle logique, il est plus facile d'identifier les différents mécanismes capables de produire les effets désirés. Comme l'explique Chen, l'objectif d'une évaluation basée sur la théorie est de fournir de l'information non seulement sur la performance d'un programme, mais aussi sur la *façon* dont il a atteint ses résultats, ainsi que sur les *raisons* qui ont mené aux effets observés.

Bien qu'on ne comprenne pas tous les mécanismes qui produisent les effets de réduction du taux de mortalité maternelle, il est clair que le contexte de la mise en œuvre des interventions liées aux SOU joue un rôle considérable. Dans les pays en développement, les SOU ne sont pas qu'une simple « recette » à appliquer pour faire diminuer les taux de mortalité maternelle. Les caractéristiques de chaque milieu (ressources, topographie, croyances, etc.) sont trop variables et trop influentes pour qu'on puisse omettre de considérer le contexte. Par contre, cette dimension et ses interactions avec l'intervention demeurent toujours peu étudiées et mal comprises.

L'analyse d'implantation du SRE dans la région de Kayes aidera à expliquer pourquoi ces interventions n'amènent pas une réduction aussi importante que prévue du taux de mortalité maternelle. Elle devrait aider à mieux comprendre les liens entre l'intervention, son contexte et les résultats observés. Plus particulièrement, l'analyse d'implantation devrait permettre de mieux connaître les effets des différentes composantes de l'intervention et de mieux définir les conditions optimales d'implantation. Cette analyse d'implantation rejoindra ainsi l'approche proposée par Patton (1997), et en particulier son concept d'analyse des composantes. Cela devrait permettre non seulement d'améliorer le SRE dans la région de Kayes, mais également d'étendre l'intervention, ou certaines de ses composantes, à d'autres milieux.

Protocole retenu pour l'évaluation

Les différentes composantes du protocole d'évaluation sont décrites au tableau 18.2.

Description du processus d'évaluation

L'évaluation a formellement démarré en septembre 2006. Nous décrivons ici les différentes étapes qui ont précédé sa mise en œuvre.

TABLEAU 18.2

Composantes de l'évaluation

Objectifs de l'évaluation	• Évaluer les effets du SRE (couverture et létalité des urgences obstétricales et des césariennes). • Comprendre les mécanismes de production de ces effets. • Déterminer les rôles du système de soins et des communautés (considérés ici comme des intrants du système) dans la production des mécanismes et des effets.
Questions posées durant l'évaluation	• Quels sont les mécanismes de production des effets du SRE? • Dans quelle mesure ces différents mécanismes permettent-ils : a) d'améliorer la couverture des urgences obstétricales et des césariennes? b) de réduire la létalité des femmes présentant une urgence obstétricale et des femmes césariées?
Méthodologie	• <i>Devis</i> : étude longitudinale avec mesures répétées pré-post pour les effets et après uniquement pour les processus et les variables associées au contexte. • <i>Sources d'informations</i> : système de suivi-évaluation et enquêtes semestrielles auprès des CSRéf. • <i>Données à recueillir</i> : <i>les intrants</i> (par exemple, nombre d'employés et qualification des membres du personnel dans les CSCom et CSRéf, équipements disponibles dans les CSRéf, pourcentage de participation financière des partenaires à la caisse de solidarité); <i>les processus</i> (par exemple, délai de transport en ambulance entre CSCom et CSRéf, délai de prise en charge au CSRéf); <i>les effets</i> (par exemple, taux de couverture des urgences obstétricales, issue de la parturiente – vivante ou décédée).
Plan de valorisation et utilisation possible des résultats	• Utiliser les données probantes pour améliorer la mise en œuvre du SRE dans la région de Kayes et dans les autres régions du Mali. • Contribuer à l'établissement d'une culture de gestion basée sur les données probantes. • Diffuser les données probantes dans la communauté scientifique.
Budget et planification	Un budget de 104 717 \$ pour une durée de trois ans (2006-2010) a été alloué pour l'évaluation de ce programme. Ce budget couvre uniquement les dépenses de recherche et il n'inclut donc pas des coûts partagés par le SRE et les dépenses directes qui y sont liées.

Phase I. Émergence du besoin d'évaluation et de formulation de la première approche (mars à juillet 2004)

À la demande de la DRS-Kayes, un des auteurs de ce chapitre a été mandaté en mars 2004 pour procéder à une revue du SRE (Fournier et Ouattara, 2004). Cette consultation visait à :

- connaître les différents besoins d'information des acteurs du système ;
- déterminer avec les responsables les principales contraintes vécues dans la mise en place et le fonctionnement du SRE depuis son implantation dans les différents districts ;
- identifier des mesures susceptibles de supprimer ces contraintes et ainsi d'améliorer le fonctionnement du système de référence-évacuation ;
- collaborer à la mise en place d'un dispositif de suivi-évaluation qui réponde aux besoins d'information des responsables et qui permette ainsi d'améliorer le fonctionnement du système de référence-évacuation et d'en évaluer l'impact.

Un atelier a permis aux responsables d'exprimer leurs besoins d'information ainsi que les différents problèmes auxquels ils faisaient face dans la mise en œuvre du SRE. Bien que présents dans tous les districts, ces problèmes ne revêtent pas tous la même importance. Par exemple :

- le taux de recouvrement des cotisations de certains partenaires (ASACO, mairies) des caisses de solidarité demeure très faible, ce qui met en péril la survie du SRE. Les acteurs du programme souhaitent disposer de données probantes qui permettraient de motiver les partenaires à cotiser selon les barèmes et les échéances fixés ;
- les comités de gestion des caisses de solidarité ne se réunissent que de façon irrégulière, et les décisions prises au cours de ces réunions ne sont pas appliquées ;
- la pénurie de ressources humaines qualifiées (médecins à compétences chirurgicales, sages-femmes, infirmières, infirmiers-anesthésistes) peut nuire à la pleine réalisation du potentiel du SRE ;
- l'accessibilité géographique demeure un problème qui se manifeste de différentes manières : les trajets en ambulance sont parfois longs et difficiles à cause des obstacles naturels (rivières, montagnes), et certaines zones deviennent inaccessibles durant la saison des pluies, ce qui réduit sensiblement la couverture potentielle du SRE à cette époque de l'année ;

- la communication par radio entre CSCom et CSRéf est parfois difficile : certains CSCom n'ont pas de radio, et les centres de santé qui ont la radio ne peuvent pas toujours l'utiliser en tout temps (à cause de piles usées ou de panneaux solaires défectueux).

Les besoins d'information qui ont été mis en évidence au cours de cet atelier avaient principalement trait aux processus et aux effets du SRE. Il est rapidement apparu que les données permettant de répondre à ces questions devaient être recueillies par un système de suivi (données de routine) et par des collectes de données plus exhaustives (qui dépassaient ce simple système de suivi) permettant de répondre à des questions précises.

La mise en place d'un système de suivi devait permettre la collecte de données sur les effets du système et sur certains indicateurs de processus. Ce système devait être simple et ne pas dépendre, à moyen terme, de l'aide extérieure. Il devait pouvoir être utilisé par les équipes de chaque district, qui deviendraient autonomes pour la collecte et l'analyse de leurs propres données, et cela dans une optique de gestion axée sur les résultats.

Des données plus exhaustives sur les facteurs susceptibles d'influer sur la mise en œuvre du programme (ressources humaines, matérielles et financières, qualité des soins, accessibilité géographique) devaient aussi être recueillies, mais hors du cadre du système de suivi, parce que leur collecte ne devait pas se faire à la même fréquence que celle des données du système de suivi et qu'elle demandait des ressources humaines et financières supplémentaires (qui pourraient éventuellement être obtenues par le moyen d'un projet de recherche).

Phase II. Première approche d'évaluation et élaboration d'un protocole de recherche (août 2004 à avril 2006)

S'appuyant sur une définition claire des besoins d'information, une première grille de collecte de données pour le système de suivi a été mise en place au cours de l'année 2004. Les responsables du SRE de chaque district ont testé la grille en colligeant les données. Ces données ont été analysées et les résultats ont été présentés au cours d'un second atelier tenu en février 2005 (Fournier et Ouattara, 2005).

Ces résultats ont été discutés par les responsables du SRE de chaque district réunis en assemblée régionale qui en ont fait des interprétations réalistes. Un des constats formulés au cours de cet atelier a été le faible

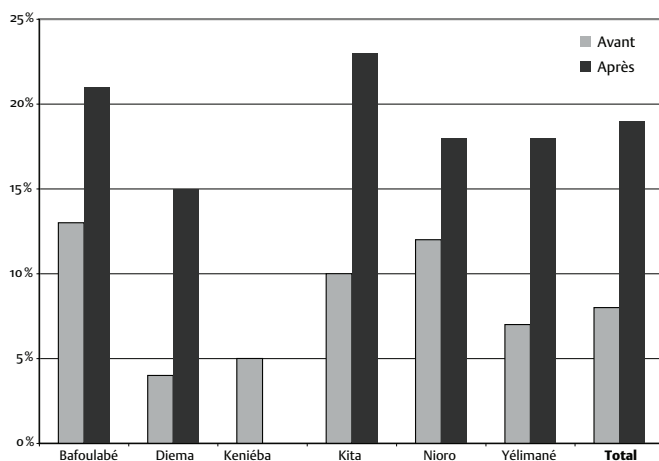
taux de réponse pour certaines variables, surtout celles qui touchaient l'analyse des processus. Des pistes de solutions ont été proposées pour y remédier. Il a été convenu qu'une deuxième version de la grille de collecte de données devait être créée. Afin de faciliter la tâche des responsables de cette collecte, des choix de réponses ont été conçus pour toutes les variables. La nouvelle grille a ensuite été proposée aux divers partenaires, qui l'ont finalement adoptée en avril 2005.

En mai et juin 2005, les responsables du SRE de six des sept districts ont effectué une collecte de données rétrospective à l'aide de la nouvelle grille. Avec la collaboration de la DRS-Kayes et d'un stagiaire de l'Université de Montréal, toutes les données sur les femmes ayant été adressées, évacuées et ayant eu recours aux SOU ont été recueillies. Cette opération avait pour but de recueillir des données exhaustives sur le SRE depuis sa mise en route dans chaque district et, lorsque c'était possible, des données sur la période qui a précédé la mise en place du système. Ces informations devaient permettre de rendre compte de la situation de la couverture en SOU et en césariennes et du taux de létalité des femmes avant la mise en place du SRE, et enfin de suivre l'évolution de ses effets depuis son implantation. Les responsables des districts ont par la suite continué à recueillir les données de façon prospective.

Cette collecte préliminaire des données dans six des sept districts de la région a montré que, avant la mise en œuvre du SRE, il y avait un écart significatif entre les districts en ce qui concerne le taux de couverture des césariennes et des urgences obstétricales, de même que leurs issues. Il y avait également, entre les districts, des différences notables dans les variations avant et après le démarrage du SRE. Par exemple :

- le taux de couverture en césariennes² avant la mise en œuvre du système varie entre 4 % et 13 %, comme le montre la figure 18.2. Les augmentations permises par la mise en œuvre du programme varient, quant à elles, entre 50 % et 400 % ;
- la létalité liée aux cas de césariennes variait entre 13 % et 19 % avant la mise en œuvre de la référence-évacuation. Les données recueillies au cours du suivi de routine montrent qu'après le début du programme ce taux a diminué de 23 à 76 %.

2. Le taux de couverture en césariennes est calculé comme le rapport entre le nombre de césariennes effectuées (numérateur) et le nombre de césariennes attendues – soit 5 % des accouchements attendus par année (dénominateur).

FIGURE 18.2**Taux de couverture en césariennes par districts (Kayes, 2002-2005)**

Ces différences entre les districts justifiaient encore davantage une analyse d'implantation du SRE, car il importait de comprendre les mécanismes de production des effets de façon générale, mais aussi pour chaque district.

Élaboration d'un protocole de recherche

Nous avons d'abord mené une réflexion concernant l'élaboration d'un projet de recherche évaluative et nous avons en vue de soumettre ce dernier à un organisme subventionnaire canadien. Par la suite, un modèle logique de cette intervention a été construit, les variables clés et les sources de données potentielles y étant déjà identifiées. En novembre 2005, une étudiante de l'Université de Montréal est allée évaluer sur le terrain la faisabilité d'une telle recherche. Elle a vérifié la source et la qualité des données et a procédé à l'opérationnalisation des variables. Il avait alors été décidé que, au cours de la réalisation du projet de recherche, ces données feraient l'objet de collectes semestrielles.

Il avait aussi été décidé que le protocole serait soumis, pour financement, au Centre de recherches pour le développement international (CRDI), dans le cadre d'un programme de recherche groupant plusieurs

équipes ouest-africaines et canadiennes. En décembre 2005, un atelier réunissant les membres de ces différentes équipes a été tenu à Ouagadougou, atelier auquel participaient également des responsables de la DRS-Kayes, eux aussi chercheurs dans ce projet de recherche. Le protocole préliminaire y a été discuté avec les participants, puis modifié et approuvé. Le programme de recherche complet a été soumis en février 2006.

Phase III. Modification et évolution de l'approche initiale (mai à octobre 2006)

Un troisième atelier réunissant les acteurs du SRE a été organisé en mai 2006. Il avait pour but de faire le point sur les données recueillies jusque-là et de dégager les résultats de l'application du programme (Fournier *et al.*, 2006).

À cette occasion, une révision radicale du système de collecte des données a été entreprise, car la grille utilisée, qui couvrait tous les besoins d'information des responsables, s'est révélée inefficace, la collecte de certaines données étant impossible ou ne donnant que de piètres résultats.

Certaines variables ont aussi été modifiées de manière à ce que le système de suivi des données de routine puisse répondre à d'autres besoins d'information surgis en cours de route, le ministère de la Santé exigeant des données sanitaires sur les SOU, la gratuité de la césarienne et la couverture des besoins obstétricaux. Le nouveau système de suivi, qui comporte une seule collecte de données, permet de répondre à toutes ces demandes.

Entre juin et octobre 2006, le système de suivi a été informatisé avec le logiciel ACCESS. L'Université de Montréal a apporté son soutien à cette démarche. Un premier atelier de formation tenu en octobre a permis de familiariser les responsables de districts (médecins et chargés du système d'informations sanitaires) avec l'entrée de données. Des modifications ont cependant été apportées à la base de données à la suite de commentaires émis par les utilisateurs. Ceux-ci peuvent maintenant entrer leurs données de manière conviviale et efficace. Cet outil informatique permet aussi de produire des rapports statistiques automatisés et de les remettre sans délai au ministère de la Santé.

Le financement demandé a été accordé, et le projet a débuté officiellement en septembre 2006.

Phase IV. Mise en œuvre effective (depuis novembre 2006)

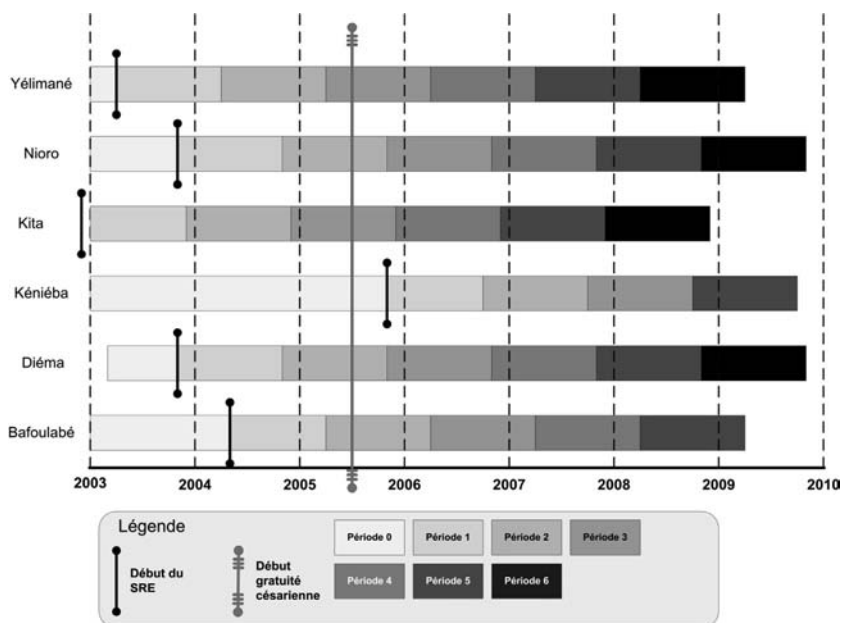
En décembre 2006, des chercheurs de l'Université de Montréal ont participé à une tournée des districts. Celle-ci avait pour but d'installer la version finale de la base de données ACCESS et de procéder à une première collecte semestrielle de données devant servir à l'analyse d'implantation du système de référence-évacuation. Ces collectes ont été effectuées à cette fréquence pendant les quatre années du projet et elles ont permis de déceler les changements survenus dans les facteurs contextuels susceptibles d'influer sur la production des effets attendus du programme.

Résultats de l'évaluation

Depuis le début de ce programme d'évaluation, plusieurs études ont été réalisées. Nous présenterons ici brièvement les résultats de la première mesure des effets (Fournier *et al.*, 2009), qui portait sur les trois années suivant le début de la mise en œuvre du SRE. Pour la commodité de notre

FIGURE 18.3

Les différentes périodes de l'évaluation



exposé, elle a été étendue ici aux trois années suivantes. L'objectif est d'estimer le taux de couverture des urgences obstétricales et la létalité des femmes présentant une complication obstétricale. Le SRE ayant commencé à des dates différentes selon les districts, les périodes ont été déterminées en fonction de la date de début dans chaque district. La figure 18.3 montre les différentes périodes couvertes par cette évaluation.

Le tableau 18.3 fournit des indications sur les activités obstétricales durant la période d'évaluation.

TABEAU 18.3**Description de la zone d'étude**

	Période 0	Période 1	Période 2	Période 3	Période 4	Période 5	Période 6
Population couverte	1 067 886	1 178 153	1 237 919	1 262 680	1 293 149	1 324 689	1 143 253
Nombre d'accouchements attendus ^a	53 394	58 908	61 896	63 134	64 657	66 234	57 163
Accouchements institutionnels ^b (%)	20,4%	27,8%	30,8%	37,3%	43,7%	47,8%	55,9%
Nombre d'urgences obstétricales traitées	455 (0,9%)	665 (1,1%)	771 (1,2%)	1200 (1,9%)	1527 (2,4%)	1660 (2,5%)	1685 (1,9%)
Interventions obstétricales majeures réalisées ^c	117 (0,22%)	270 (0,46%)	373 (0,60%)	482 (0,79%)	586 (0,91%)	695 (1,05%)	658 (1,16%)
Indications maternelles absolues (IMA) ^d	66 (0,12%)	131 (0,22%)	202 (0,37%)	274 (0,43%)	303 (0,47%)	377 (0,57%)	329 (0,58%)
Non IMA	41 (0,08%)	133 (0,23%)	166 (0,27%)	208 (0,33%)	283 (0,44%)	318 (0,48%)	329 (0,58%)
Inconnu	10 (0,02%)	6 (0,01%)	5 (0,01%)	0	0	0	0

a. Calculé comme 5 % de la population. Ce nombre a été retenu pour calculer les pourcentages.

b. Inclut les accouchements dans les CSCOM (1^{er} niveau) et les CSRéf (2^e niveau).

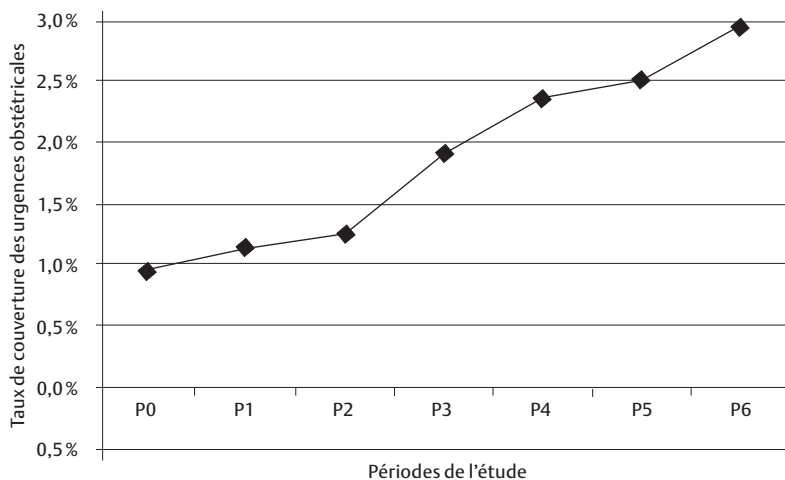
c. Inclut la césarienne, l'hystérectomie, la laparotomie pour rupture utérine, la version par manœuvre interne et la craniotomie/embryotomie.

d. Inclut les hémorragies ante-partum sévères et post-partum incoercibles, la prérupture et rupture utérine, la disproportion fœto-pelvienne et la présentation frontale et transverse.

La figure 18.4 montre l'évolution du taux de couverture des urgences obstétricales durant la période sous revue.

FIGURE 18.4

Taux de couverture des urgences obstétricales selon la période



Pour la période sous revue, on peut constater qu'entre la première année avant le début du SRE et les deux années après, la croissance annuelle a été en moyenne de 15 %. Pour la troisième à la sixième année après, cette croissance a été en moyenne de 34 % par an.

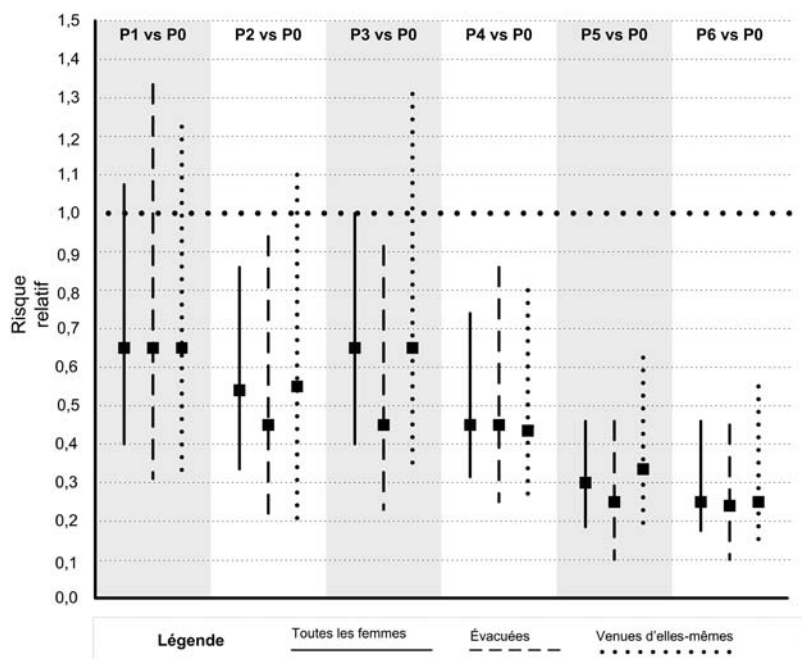
La létalité des urgences obstétricales

La figure 18.5 montre les risques relatifs (RR) des femmes qui ont présenté une complication obstétricale. Ces risques relatifs permettent de suivre l'évolution du risque de décès chaque année à partir du début du SRE en utilisant toujours la période avant SRE comme référence. Deux sous-groupes ont été constitués : les femmes évacuées, c'est-à-dire celles qui ont bénéficié des trois composantes du programme et celles qui sont venues d'elles-mêmes, c'est-à-dire qui n'ont pas utilisé l'ambulance du SRE. On s'attend à ce que les bénéfices de l'intervention soient plus importants

pour le premier groupe. Les risques relatifs (voir l'encadré 1) ont été ajustés par plusieurs variables connues de façon qu'ils deviennent des facteurs de risque : l'âge, le type de complication, l'antécédent de césarienne, l'accessibilité géographique et le type de traitement reçu (césarienne, transfusion sanguine).

FIGURE 18.5

Risques relatifs (RR) des femmes qui ont présenté une complication obstétricale



ENCADRÉ 1

Définition du risque relatif

Taux de létalité après l'intervention rapporté au taux de létalité avant l'intervention. Ce rapport est généralement considéré comme une mesure de la force de l'association existant entre l'exposition et le résultat. Il répond à la question : combien de fois les sujets exposés ont-ils plus de chances que les sujets non exposés de décéder. Un risque relatif inférieur à 1 signifie que les sujets exposés à l'intervention ont moins de chances de décéder que les ceux qui ne l'ont pas été (adapté de Landrison et Deklahaye, 1995).

À partir de la quatrième année après le SRE, le risque relatif se situe à 0,43 pour les femmes venues d'elles-mêmes, 0,47 pour les femmes évacuées et 0,47 pour les deux groupes réunis. Cela signifie que le risque de décès a été divisé par plus de deux pour toutes les utilisatrices, comparativement à la situation initiale. Cette diminution a été encore plus rapide pour les femmes qui ont été évacuées, puisque ce résultat a été obtenu dès la deuxième année. Des analyses plus détaillées ont montré que 40 % de la réduction du risque de mortalité observée chez les femmes évacuées était attribuable à la réduction de la mortalité par hémorragie. Cela et le fait que le bénéfice de l'intervention a été plus rapide chez les femmes qui ont bénéficié de toutes ses composantes accréditent la thèse selon laquelle les effets observés sont dus au SRE. À partir de la cinquième année, on observe une nouvelle baisse pour tous les groupes de femmes, leur risque de décès est trois à quatre fois moins élevé qu'avant la mise en place du SRE (risque relatif de 0,30 pour toutes les femmes, de 0,33 pour les femmes venues d'elles-mêmes et de 0,23 pour les femmes évacuées). Le résultat est le même dans la sixième année.

Discussion

Même si l'évaluation de ce programme est toujours en cours, nous pouvons déjà tirer plusieurs leçons de notre expérience. Premièrement, le fait que les besoins d'information ont été exprimés, définis et ajustés par les acteurs du programme a grandement aidé à franchir les différentes étapes de l'évaluation. Il a entre autres permis de concilier les intérêts des différents acteurs et de les amener à poursuivre un objectif commun : mieux comprendre le système de référence-évacuation, son fonctionnement et la production de ses effets.

La DRS-Kayes et l'Université de Montréal ont ainsi pu bâtir un partenariat basé sur la mise en commun des ressources financières, humaines et logistiques pour la réalisation des activités (collecte et analyse des données), afin de faciliter le travail dans des conditions contextuelles et logistiques parfois complexes. Il arrive que les chercheurs de l'Université de Montréal qui participent au projet de recherche appuient la DRS-Kayes dans des dossiers liés au SRE, en fournissant par exemple les analyses de données devant figurer dans les rapports destinés à l'administration centrale ou aux équipes sociosanitaires des districts.

L'importance d'informer de façon continue la Direction nationale de la santé au sujet de l'évolution du système de collecte est une deuxième leçon que nous avons tirée de notre expérience. Les mises à jour ont été effectuées sur une base régulière par l'équipe de la DRS-Kayes.

Rappelons qu'un tel suivi de l'évolution des soins obstétricaux d'urgence dans les pays en développement a rarement été fait au cours des dernières années, ce qui explique, en partie, les faibles progrès réalisés dans la réduction de la mortalité maternelle. Un exemple probant d'intervention dans un pays de l'Afrique sub-saharienne pourra aider à démontrer que le suivi des SOU est possible.

Forces et faiblesses du processus

Le partenariat établi avec la DRS-Kayes demeure le point central de cette évaluation. Il a permis de réaliser le programme dans les délais fixés et chaque partenaire a pu satisfaire ses besoins d'information.

Par ailleurs, dans un contexte où la DRS-Kayes a une multitude de dossiers sanitaires à traiter et à gérer, et où les priorités sont souvent imposées par le pouvoir central, on peut se demander s'il convient que la DRS-Kayes travaille dans une optique de qualité sur un dossier commun avec l'Université de Montréal. On pourrait craindre que cette démarche ait pour effet de concentrer les efforts sur le SRE au détriment des autres enjeux du secteur de la santé qui se posent dans cette région. Cette crainte ne paraît cependant pas fondée, car la démarche d'évaluation et le cheminement effectué avec la DRS-Kayes et les sept équipes sociosanitaires de district ont permis de créer une culture de prise de décision basée sur les données probantes, le personnel consultant et analysant maintenant les données recueillies, et fondant sa planification sur les résultats obtenus.

Cela dit, la pérennité d'une telle démarche n'est pas assurée, car, comme nous l'avons mentionné, l'Université de Montréal a déjà financé plusieurs tournées de collecte de données dans les districts et fourni des ressources humaines pour le traitement et l'analyse de ces données. On ne peut que souhaiter que les activités se maintiennent après que le projet de recherche aura pris fin.

Principaux enjeux lors de la mise en œuvre de l'évaluation

La qualité des données recueillies dans chacun des districts a été un enjeu majeur dans la mise en œuvre de l'évaluation, car, même si elle s'est améliorée au cours de la dernière année, elle demeure néanmoins un sujet important de préoccupation et demande un effort constant et prolongé. Ce point pourrait constituer un sujet de désaccord entre les chercheurs qui veulent maximiser la qualité des données et les acteurs sur le terrain, pour qui les soins obstétricaux et le système de référence-évacuation ne constituent qu'une problématique parmi d'autres.

Principaux résultats de l'évaluation

Les résultats de la mesure des effets du SRE présentés dans ce chapitre montrent que les effets attendus sont observés : la couverture en soins obstétricaux augmente et le taux de létalité diminue. Ce devis observationnel ne permet pas d'attribution causale univoque (voir chapitre 6), mais une analyse plus détaillée de l'évolution de la létalité suggère que les effets observés peuvent raisonnablement être attribués au SRE. Ces résultats sont encourageants pour les responsables sanitaires et sociaux de la région de Kayes et pour le projet canadien, car ils montrent que leurs efforts ont été fructueux. Ils le sont également pour les autorités nationales, puisque le SRE est un programme essentiel. Il faut néanmoins éviter de se montrer trop optimiste en ce qui concerne la généralisation de ce programme à d'autres régions du Mali ou à d'autres pays. Il y a en effet deux facteurs contextuels qui ont été déterminants dans le succès du SRE à Kayes. D'une part, la région a bénéficié pendant toute la période d'un appui technique et financier du Canada et, d'autre part, les besoins étaient si importants qu'il était assez aisé d'obtenir des résultats positifs. En effet, plus les besoins sont grands, plus ils sont faciles à combler et, à mesure qu'ils sont satisfaits, la tâche devient de plus en plus complexe. D'ailleurs on observe que le taux de couverture en urgences obstétricales continue de croître et que la létalité ne diminue plus. Le niveau actuel de couverture en urgences obstétricales (2,9 %) est encore loin des taux optimaux que l'on estime à 10 % à 15 %. Le fait que la létalité ne diminue plus pourrait signifier que l'intervention ne rejoint pas de nouvelles strates de population auxquelles l'intervention bénéficierait. Des analyses plus détaillées

de la progression de l'intervention dans différentes strates de la population permettront de vérifier cette hypothèse.

Conclusion

Les premiers résultats de l'analyse des effets du SRE dans la région de Kayes démontrent qu'il est possible de réduire la létalité des urgences obstétricales, et cela, en quelques années seulement. L'analyse d'implantation permettra de comprendre les mécanismes qui produisent ces effets, afin d'améliorer le potentiel d'action du SRE et, éventuellement, d'assurer sa généralisation à d'autres contextes.

Pour aller plus loin : lectures complémentaires sur le Mali

AUDIBERT, M. et E. de ROODENBEKE (2005), *Utilisation des services de santé de premier niveau au Mali : analyse de la situation et perspectives*, Banque mondiale, Région Afrique, Département du développement humain.

MINISTÈRE DE LA SANTÉ ET SECRÉTARIAT GÉNÉRAL DE LA SANTÉ ET PAUVRETÉ (septembre 2004), *La problématique de la santé et de la pauvreté au Mali. Analyse des indicateurs de santé et population dans le cadre de la stratégie nationale de lutte contre la pauvreté au Mali*, Banque mondiale, Région Afrique.

D'autres aspects de l'évaluation dans ce programme

DOGBA, M. *et al.* (2011), Mother and newborn survival according to point of entry and type of human resources in a maternal referral system in Kayes (Mali), *Reproductive Health*, 8, 13.

DOGBA, M., P. FOURNIER et S. BERTHE-CISSE (2012), Qualification of staff, organization of services, and management of pregnant women in rural settings: The case of Diema and Kayes districts (Mali), *ISRN Obstetrics and Gynecology*, vol. 2012, Article ID 649412.

PIRKLE, C. M. *et al.* (2011), Emergency obstetrical complications in a rural African setting (Kayes, Mali) : The link between travel time and in-hospital maternal mortality, *Maternal and Child Health Journal*, 15(7), 1081-1087.

19

Une évaluation de l'efficacité d'un programme de formation

Diane Berthelette et Luc Desnoyers

Le contexte

C'est en participant à des activités de formation syndicale en santé du travail que nous nous sommes interrogés sur le problème de leur évaluation scientifique (Desnoyers, 1986). Mais pareil projet de recherche ne pouvait se concrétiser que dans un contexte particulier. Il fallait conclure un accord préalable avec une centrale syndicale, choisir d'évaluer une activité offerte sur une base suffisamment fréquente, et développer un cadre théorique et des outils d'évaluation pertinents.

Nous nous sommes intéressés au programme de formation de base en santé et en sécurité du travail élaboré par une des centrales syndicales québécoises. L'objectif ultime de ce programme dispensé depuis 1976 était d'améliorer l'action syndicale des travailleurs dans leurs milieux de travail à l'égard de la prévention primaire des lésions professionnelles. Nous avons proposé au directeur du service d'éducation de cette centrale d'analyser les effets attendus du programme en question. Nous avons cependant insisté sur la nécessité de comprendre sa théorie sous-jacente avant de procéder à l'évaluation de son efficacité, en raison des limites importantes

de l'approche de la « boîte noire » (Chen et Rossi, 1983). Grâce à son accord, nous avons obtenu la collaboration du permanent syndical avec lequel nous donnions nos cours.

Notre recherche a été subventionnée par l'Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST). Cet organisme paritaire est essentiellement financé par les cotisations versées par les employeurs du Québec à la Commission de la santé et de la sécurité du travail (CSST).

La phase I de l'étude a consisté à décrire la théorie sous-jacente à ce programme de formation, à en évaluer la plausibilité en fonction des connaissances scientifiques disponibles et à estimer l'écart entre le programme prescrit et celui qui était effectivement dispensé (Berthelette *et al.*, 1998). Dans la phase II, une analyse des effets, il s'est agi d'évaluer l'efficacité de l'utilisation du programme.

La théorie sous-jacente au programme

Le programme de formation de base en santé et en sécurité du travail s'adresse à des membres de comités de santé et de sécurité du travail, à des délégués syndicaux, à des officiers ou à de simples syndiqués. Il est offert régulièrement par des formateurs qui sont eux-mêmes des militants syndicaux formés par la centrale. Les novices dispensent le cours sous la supervision de formateurs supérieurs.

Selon Blondin (1980), qui est le principal concepteur de ce programme, celui-ci doit suivre les trois étapes suivantes :

1. favoriser la « prise de conscience du nombre dramatique d'accidents de travail et de la détérioration souvent rapide de la santé des travailleurs comme conséquence de la présence peu contrôlée de tueurs silencieux qui hypothèquent gravement leur santé » ;
2. transmettre des « informations sur les lois dont le but officiel est la protection de la santé » ;
3. traiter des « diverses dimensions des moyens d'action syndicale... pour protéger [la] santé, assurer [l']intégrité physique... ».

Ces trois étapes correspondent aux trois parties du programme de formation présentées dans les cahiers du formateur et du participant, soit les accidents de travail et atteintes à la santé, les droits des travailleurs

et des travailleuses en matière de santé et de sécurité, et l'action syndicale.

Les formateurs et les participants utilisent les cahiers élaborés par la centrale et en suivent le contenu de manière séquentielle. Il existe toutefois un écart entre les durées de formation prescrites dans les cahiers et celles que nous avons pu observer. La durée de formation totale prescrite est de 15 heures, alors que celle que nous avons mesurée a été de 12 ½ heures. De plus, les formateurs consacrent 1,5 fois plus de temps à la première partie du cours que ce qui est prescrit, 0,6 de la durée prévue pour la partie portant sur les droits des travailleurs, et 0,46 du temps de formation demandé pour la partie portant sur l'action syndicale. Ces écarts seraient attribuables aux quatre dimensions suivantes de la théorie sous-jacente au programme :

- l'objectif ultime du programme est d'amener les participants à mobiliser les travailleurs pour qu'ils fassent pression sur les dirigeants de l'entreprise afin d'éliminer les sources de risque de lésions professionnelles ;
- pour atteindre cet objectif, le programme doit au préalable amener les participants à attribuer l'ensemble des lésions professionnelles aux conditions de travail, et non plus à une « faute » des travailleurs ;
- les conditions de travail relèvent de la responsabilité des employeurs ;
- il est inutile de poursuivre la formation si les objectifs du premier bloc du cours ne sont pas atteints.

Les activités de formation sont de quatre types : des exposés théoriques dispensés par le formateur (environ 35 % du cours), des plénières au cours desquelles le formateur demande des informations aux participants (environ 5 % du cours), des discussions en sous-groupes des cas décrits dans le matériel pédagogique (environ 40 % du cours), et des plénières au cours desquelles un représentant de chacun des sous-groupes présente les résultats des discussions (environ 20 % du cours). D'après les travaux de Freire (1983) et de Blondin (1980), la théorie sous-jacente au programme implique que les formateurs demandent constamment aux participants d'exprimer leurs connaissances et leurs perceptions, d'une part, sur leurs conditions de travail et sur leur santé, et, d'autre part, sur les comportements des personnes concernées par les problématiques de santé et de sécurité du travail. Ils doivent également leur demander de dire ce

qu'ils pensent des effets de leur travail sur leur santé. Le formateur doit interagir avec les participants pour les aider à renforcer ou à modifier leurs opinions.

Protocole

L'objectif de l'évaluation était d'analyser les effets intermédiaires du programme de formation. Il nous apparaissait plus utile de centrer notre évaluation sur ceux-ci que de mettre d'emblée l'accent sur les effets ultimes du programme, puisqu'il s'agissait de la première recherche évaluative, dont les résultats devaient être publiés, sur un programme de formation syndicale en santé et en sécurité du travail. Les effets ultimes dépendent par ailleurs de trop de facteurs externes à l'activité de formation pour qu'il soit pertinent de les évaluer ici.

Dans le présent chapitre, nous nous centrons uniquement sur l'un des thèmes du programme, soit les accidents de travail. Afin d'optimiser la validité des instruments de mesure des effets attendus relatifs à ce thème, nous avons fait appel à une approche inductive pour choisir nos cadres conceptuels, c'est-à-dire que nous avons utilisé ceux qui correspondaient le mieux à la nature des effets que nous avons identifiés au cours de la phase I de notre étude. Il s'agit de la taxonomie des objectifs pédagogiques (Bloom *et al.*, 1969) et de la théorie de l'attribution (Dubois, 1996). Nous aurions pu être tentés, comme d'autres évaluateurs, d'avoir recours au cadre conceptuel de Donald L. Kirkpatrick (1996) en raison de la renommée de cet auteur. Ce dernier propose d'évaluer les quatre catégories suivantes d'effets des programmes de formation : la réaction des participants au programme ; l'apprentissage qui cible les changements dans les attitudes, les connaissances et les habiletés qui devraient survenir après la formation ; les comportements des personnes formées ; et enfin les résultats qui se traduisent par des changements de nature organisationnelle, comme la productivité et la qualité des activités de production. Mais ce cadre d'analyse comporte selon nous deux limites importantes. D'une part, il ne tient pas compte des données probantes produites par des experts en évaluation tels Chen et Rossi. Depuis près de deux décennies, leurs études montrent ainsi que les effets des programmes à évaluer sont liés à la théorie des programmes en question. Pour sa part, Kirkpatrick ne mentionne pas cet aspect. Au contraire, il recommande l'utilisation de son cadre pour

tout type de programme de formation implanté en entreprise. D'autre part, les catégories d'effets qu'il propose ne sont pas mutuellement exclusives, des études en psycho-sociologie ayant démontré que des changements de comportements peuvent être associés à l'apprentissage.

La taxonomie des objectifs pédagogiques distingue six niveaux d'apprentissage, subdivisés en sous-catégories et ordonnés de manière hiérarchique en fonction de leur degré d'abstraction et de complexité, soit : l'acquisition des connaissances, leur compréhension, leur application, leur analyse, leur synthèse et leur évaluation. Ce cadre conceptuel nous a permis de cerner deux des objectifs d'apprentissage, soit la compréhension des facteurs qui favorisent la survenue des accidents et l'application de connaissances à l'égard de leur prévention. La compréhension consiste à connaître l'information qui est communiquée et à savoir s'en servir, sans nécessairement établir de liens avec d'autres informations ni en saisir toute la portée. Quant à l'application des connaissances, elle est définie comme « l'utilisation des représentations abstraites dans des cas particuliers et concrets ».

L'attribution est définie comme « une explication, un jugement porté, une inférence faite *a posteriori* sur un événement précis » (Dubois, 1996). Elle porte sur les lieux de causalité de cet événement, qui peuvent être d'ordre interne (les causes étant attribuées à des caractéristiques de la personne impliquée dans l'événement) ou externe (les causes étant attribuées à des facteurs d'origine environnementale). Ce concept nous a permis de qualifier l'un des effets attendus de la formation, soit l'habileté des personnes formées à expliquer la fréquence des accidents de travail par des facteurs environnementaux, plutôt que de les attribuer au fait que les travailleurs victimes des accidents adoptent des comportements à risque. Il s'agit en fait de la pierre angulaire de la théorie sous-jacente au programme élaboré par la centrale.

L'analyse qualitative des données de la phase I a permis de dégager les objectifs intermédiaires suivants poursuivis à l'égard du thème des accidents de travail : comprendre que les accidents de travail ont plusieurs causes, être capable de déterminer quelles sont ces causes, être capable de trouver des mesures permettant d'éviter que des accidents se produisent et pouvoir attribuer les accidents de travail à des risques environnementaux dont les employeurs sont responsables. Il s'agit là des effets que notre recherche avait pour but d'analyser.

Nous avons utilisé un devis prétest/post-test avec groupe témoin non équivalent, car il était impossible d'assigner aléatoirement les sujets entre le groupe participant¹ au programme et le groupe témoin, la mise en œuvre du programme et le recrutement des participants étant sous le contrôle de la centrale syndicale, de ses instances régionales et des syndicats qui la composent. Le choix d'un tel devis nous permettait de réduire au minimum le nombre de biais susceptibles d'altérer la validité interne de nos inférences.

Afin d'optimiser l'utilisation de nos résultats par la centrale, nous avons invité le permanent syndical responsable des cours de santé et de sécurité du travail à prendre part à nos décisions, et l'avons régulièrement tenu informé de l'évolution du projet. Nous avons également présenté nos résultats finaux au directeur du service d'éducation de la centrale et aux membres de son équipe.

L'IRSST nous a d'abord octroyé une subvention de 90 558 \$ pour trois ans, mais nous avons obtenu une prolongation de un an et un budget supplémentaire de 25 000 \$ en raison de la complexité du projet.

Description du processus d'évaluation

La population étudiée était composée des membres des syndicats locaux de la centrale syndicale, et nous avons choisi d'utiliser une méthode d'échantillonnage par grappes pour recruter les participants. Nous sommes d'abord entrés en contact avec les formateurs qui dispensaient des cours durant la période prévue pour la collecte de données. Nous leur avons expliqué les objectifs de l'étude et les méthodes utilisées, et ils ont tous accepté d'y participer. Quant aux personnes qui suivaient les cours, nous leur avons soumis un formulaire de consentement dès le premier cours. Le prétest a été administré au début de la première journée de formation, alors que le post-test l'a été à la fin de la dernière journée.

L'échantillon du groupe des participants était composé de 65 délégués syndicaux ayant participé à quatre sessions de formation dispensées entre les mois de février et de novembre 1999. Ces sessions regroupaient en moyenne 18 personnes et étaient dispensées par deux formateurs sur une

1. On aura compris que la notion d'exposition à un programme de formation est incongrue, d'où notre choix d'appeler « groupe participant » ce qui est formellement nommé « groupe exposé ».

période de trois journées consécutives, conformément au programme prescrit.

Les personnes du groupe témoin ont été recrutées parmi les membres des syndicats locaux participant à d'autres programmes de formation que celui qui faisait l'objet de notre étude, entre les mois de novembre 1997 et février 2000 ($n = 60$). Elles participaient à l'un des six programmes de formation portant sur les thèmes suivants: le bruit, l'organisation du travail (deux groupes) et le délégué syndical (trois groupes). Les activités de ces programmes étaient réparties sur trois jours, comme celles du programme étudié.

Afin de construire un instrument de mesure reflétant adéquatement l'importance relative des effets attendus du programme qui devaient être mesurés, nous avons créé un tableau de spécification des objectifs d'apprentissage, de leur contenu notionnel et de leur importance relative (Tousignant et Morissette, 1990). Nous avons par la suite élaboré sur cette base un questionnaire auto-administré comportant des questions relatives aux effets attendus du programme et aux caractéristiques des participants. Ces dernières devaient nous permettre de contrôler d'éventuels biais de sélection. Nous avons enfin soumis ce questionnaire au permanent syndical avec lequel nous collaborions, afin d'en valider le contenu.

Nous avons mesuré l'ensemble des variables dépendantes qui font l'objet du présent chapitre à l'aide du cas suivant, qui était comparable à ceux des cahiers de formation :

Diane est technicienne de laboratoire. Elle porte des moyens individuels de protection (MIP) dans le cadre de ses activités. Il s'agit d'un sarrau, de gants, de lunettes, d'un chapeau et d'un masque respiratoire. Elle travaille dans un environnement dont la température varie entre 25 et 35 degrés Celsius. Cette température élevée est due essentiellement à la proximité d'appareils de stérilisation. Avec la chaleur, le port de l'équipement devient particulièrement irritant et il arrive que Diane l'enlève. Un jour, dans cette situation, Diane est brûlée au deuxième degré par les gaz qui se sont échappés de l'appareil de stérilisation au moment où elle l'a ouvert.

Les trois questions suivantes ont été utilisées :

1. Quelles sont la ou les causes de cet accident ?
2. Comment cet accident aurait-il pu être évité ?
3. Qui est responsable de cet accident ?

La première question de notre instrument de mesure était ouverte. Elle visait d'abord à mesurer la compréhension des participants quant au phénomène de l'accident, mais il s'agissait également d'une façon de mesurer l'attribution des causes de l'accident. Nous avons répertorié l'ensemble des réponses fournies par les participants, et les chercheurs responsables du projet ont procédé à une évaluation normative de la plausibilité théorique et technique de chacune des réponses. Les réponses plausibles à la première question ont par la suite été codifiées à nouveau dans les deux catégories suivantes : causes individuelles et causes environnementales. Enfin, pour chaque participant, nous avons calculé le nombre de réponses appartenant à chacune de ces catégories. Il s'agit des deux premières variables dépendantes, soit le nombre de causes individuelles et le nombre de causes environnementales plausibles de l'accident décrit dans ce cas énumérées par les participants.

La deuxième question était aussi ouverte. Elle visait à mesurer l'application des connaissances acquises à l'égard de la prévention des accidents de travail. Les réponses plausibles à la deuxième question ont été codifiées selon les deux catégories suivantes : les actions individuelles plausibles de la part de la travailleuse et les actions plausibles de la part de l'employeur. Il s'agit de nos troisième et quatrième variables dépendantes, soit le nombre d'actions individuelles et organisationnelles plausibles qui auraient pu être menées pour prévenir l'accident décrit dans ce cas.

La troisième question comportait les deux consignes suivantes, dont les réponses étaient mesurées à l'aide d'échelles visuelles analogues :

1. indiquez le pourcentage de responsabilité de l'employeur dans cet accident ;
2. indiquez le pourcentage de responsabilité de la travailleuse dans cet accident.

Les pourcentages de responsabilité de l'employeur et de la travailleuse sont les deux dernières variables dépendantes de notre étude. Leur valeur peut varier entre 0 et 10.

Les interactions entre la sélection, d'une part, et l'accoutumance au test, la maturation et l'histoire, d'autre part, sont des biais susceptibles de menacer la validité interne d'une étude basée sur un devis comme le nôtre (Shadish *et al.*, 2002). En effet, l'absence d'une assignation aléatoire des sujets entre le groupe participant et le groupe témoin peut faire en sorte

que les groupes présentent des caractéristiques différentes dès le prétest, susceptibles de produire un écart entre leurs résultats respectifs lors du post-test. Afin d'éviter qu'un tel biais de sélection ne menace la validité interne de nos inférences, nous avons inclus les résultats du prétest pour chacune des variables dépendantes (les mesures d'effets) dans nos analyses multivariées à titre de variables de contrôle. Nous avons également contrôlé la présence possible de biais d'interaction entre la sélection et des biais d'accoutumance au test, de maturation ou d'histoire. Ce faisant, nous avons pu estimer l'effet net de la formation, soit le degré d'apprentissage des sujets entre la période qui précédait la formation (prétest) et celle qui la suivait (post-test).

Un *biais d'accoutumance au test* apparaît lorsque les réponses au post-test sont altérées par l'utilisation de l'instrument de mesure des effets lors du prétest. Il était en effet possible que la connaissance des questions relatives aux effets attendus du programme acquise lors de l'administration du prétest entraîne chez les participants une réflexion pouvant modifier les réponses données lors du post-test, et cela, plus particulièrement chez les sujets qui possédaient déjà des compétences en santé et en sécurité du travail au moment du prétest. C'est pourquoi nous avons introduit les trois variables suivantes, relatives aux compétences en santé et en sécurité du travail acquises avant la participation au cours, à titre de variables de contrôle: le nombre de mois d'expérience au sein d'un comité de santé et de sécurité du travail, le nombre de mois d'expérience à titre de délégué syndical et la participation antérieure à un cours de formation syndicale en santé et en sécurité du travail. Nous avons mesuré les deux premières variables, qui sont de nature continue, à l'aide de questions ouvertes. Nous avons attribué une valeur nulle à ces deux variables lorsque les participants ne possédaient aucune expérience. Pour la troisième variable, qui est dichotomique, nous avons demandé aux participants de nommer les programmes de formation qu'ils avaient suivis à l'aide d'une liste de 23 programmes répertoriés par la centrale. L'exposition antérieure à l'un ou l'autre de trois programmes pertinents (Actions en prévention, Organisation du travail et Plaideurs à la Commission de la santé et de la sécurité du travail) faisait attribuer un code signifiant la présence d'une formation en santé et en sécurité du travail.

Un *biais de maturation* altère la validité interne des résultats lorsque les sujets subissent des changements liés au temps écoulé entre le prétest

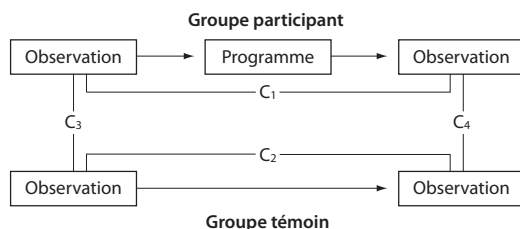
et le post-test, changements susceptibles d'altérer leurs résultats du post-test. Un *biais d'histoire* est présent lorsque les sujets sont exposés, entre le prétest et le post-test, à une autre intervention qui poursuit des objectifs similaires à celle qui fait l'objet de la recherche évaluative.

Nous avons administré les questionnaires du post-test lors de la dernière journée de formation, soit trois jours après le début de la formation, et cela dans les deux groupes afin de réduire au minimum la probabilité d'interaction entre d'éventuels biais de sélection, de maturation et d'histoire.

Nous avons effectué des analyses descriptives des variables dépendantes et de contrôle, de même que des analyses bivariées pour comparer nos groupes. Dans un premier temps, nous avons vérifié s'il existait des différences significatives entre le groupe participant et le groupe témoin à l'égard des variables de contrôle en faisant appel à des tests de t (expérience dans un comité de SST et à titre de délégué syndical) et à des tests de chi carré (nombre de sujets ayant une expérience dans un comité de SST, ayant occupé une fonction de délégué syndical et ayant reçu une formation antérieure en SST). Comme le montre la figure 19.1, nous avons évalué les écarts entre les groupes pour chacune des variables dépendantes à l'aide de tests de t :

1. prétest vs post-test
 - a) pour le groupe participant (C₁)
 - b) pour le groupe témoin (C₂) ;
2. prétests des groupes participant et témoin (C₃) ;
3. post-tests des groupes participant et témoin (C₄).

FIGURE 19.1



En incluant les variables de contrôle dans nos analyses multivariées, nous pouvions non seulement estimer les effets de ces variables, mais aussi

les contrôler, de manière à optimiser la validité interne de nos inférences. Ce faisant, nos résultats permettaient d'estimer l'effet net de la participation au programme de formation sur les effets attendus en matière d'accidents de travail. La participation au programme, mesurée de manière dichotomique selon l'appartenance des sujets au groupe participant ou au groupe témoin, constituait la variable indépendante. Les résultats obtenus par les sujets lors du post-test, concernant les effets attendus du programme, correspondaient aux six variables dépendantes.

Nos analyses descriptives ont révélé que les quatre premières variables dépendantes suivaient une distribution de Poisson. Puisque les prémisses de la régression multiple linéaire n'étaient pas respectées, nous avons regroupé les réponses en deux catégories, pour les analyses finales, soit la présence ou l'absence de réponses plausibles. Puis nous avons soumis nos données à des analyses de régression logistique multiple. Nous avons utilisé la méthode de sélection des variables progressive hiérarchique. La sélection des variables a été réalisée à l'aide du ratio des logarithmes naturels de vraisemblance (*likelihood ratio*). Les deux méthodes qui nous ont permis d'évaluer le degré d'ajustement du modèle sont la comparaison des valeurs prédites et des valeurs observées, et le rapport de vraisemblance.

Les deux dernières variables dépendantes ont été soumises à des analyses de régression linéaire multiple. Nos deux modèles de régression multiple linéaire en respectent les prémisses, comme l'ont révélé les analyses résiduelles et les analyses des corrélations entre les variables indépendantes et de contrôle. Le coefficient de corrélation le plus élevé était de 0,415 ($p = 0,000$) ; il concernait le nombre de mois d'expérience en SST et l'ancienneté dans une fonction de délégué syndical. Nous avons fait appel à la méthode de sélection des variables progressive hiérarchique, les variables de contrôle étant introduites avant la variable indépendante.

Résultats

Les taux de réponse des groupes participant et témoin sont respectivement de 80 % et de 90 %. Toutefois, après avoir retiré les questionnaires qui ne contenaient pas l'ensemble des données utiles aux analyses, les taux de réponse chutent à 65 % (groupe participant) et à 70 % (groupe témoin), chacun des groupes étant composé de 42 sujets.

Trois fois plus de personnes du groupe participant ($n = 30$) que du groupe témoin ($n = 10$) ont été membres d'un comité de santé et de sécurité du travail ($p = 0,000$). Le nombre moyen de mois d'expérience dans un tel comité était de 7,3 mois ($s = 9,6$) chez les sujets du groupe participant et de 11,1 mois ($s = 22,1$) chez les sujets du groupe témoin. Cet écart n'est pas statistiquement significatif ($p = 0,456$).

Le nombre de sujets possédant une expérience de délégué syndical était de 23 chez le groupe participant et de 34 chez les témoins. Cette différence est statistiquement significative ($p = 0,009$). Le nombre moyen de mois d'expérience dans cette fonction s'élève à 11,1 mois ($s = 22,1$) chez le groupe participant et à 36,4 mois ($s = 66$) chez le groupe témoin ($p = 0,021$).

Par ailleurs, les pourcentages de personnes ayant suivi d'autres programmes de formation en SST que le programme étudié sont de 2 % pour le groupe participant et de 33 % pour le groupe témoin. Cette proportion élevée dans le groupe témoin n'est pas surprenante, puisque le processus de trois des six programmes dans lesquels nous les avons recrutées portait en partie sur des aspects liés à la santé ou à la sécurité du travail. Nous n'avons pu effectuer de test de chi carré pour cette variable, puisqu'une cellule contenait moins de cinq sujets. L'importance de l'écart observé entre les groupes à l'égard de formations antérieures en SST semble indiquer la présence d'un biais de sélection entre les groupes.

Les analyses bivariées présentées dans les tableaux 19.1 et 19.2 révèlent la présence d'une seule différence statistiquement significative entre les groupes participant et témoin en ce qui concerne les résultats au prétest (C3 de la figure 19.1). Il s'agit de l'appréciation de la responsabilité de l'employeur dans l'accident décrit dans le cas soumis aux sujets. Le groupe participant accorde au départ une plus faible responsabilité à l'employeur que ne le fait le groupe témoin ($5,45 \pm 3$ contre $6,82 \pm 3,02$). La présence d'un tel écart significatif indique un biais de sélection, pour cette variable, mais que nous avons contrôlé dans les analyses finales par l'inclusion des résultats du prétest à titre de variables de contrôle.

TABEAU 19.1**Résultats descriptifs et analyses comparatives des résultats prétest et post-test (n = 84)**

Variable	Prétest			Post-test		
	Témoins $\bar{x} \pm s$	C ₃ p	Participants $\bar{x} \pm s$	Témoins $\bar{x} \pm s$	C ₄ p	Participants $\bar{x} \pm s$
Causes individuelles	0,278 ± 0,452	0,134	0,423 ± 0,537	0,286 ± 0,486	0,000*	0,038 ± 0,194
Causes environnementales	0,944 ± 0,787	0,122	0,712 ± 0,750	0,679 ± 0,716	0,003*	1,15 ± 0,916
Actions individuelles	0,122 ± 0,298	0,204	0,232 ± 0,396	0,183 ± 0,382	0,006*	0,008 ± 0,053
Actions organisationnelles	1 ± 0,911	0,148	0,731 ± 0,992	0,714 ± 0,803	0,000*	1,5 ± 0,828
% responsabilité employeur	6,82 ± 3,02	0,027*	5,45 ± 3	7,59 ± 2,67	0,000*	9,64 ± 1,55
% responsabilité travailleuse	4,8 ± 3,27	0,146	5,71 ± 2,84	3,22 ± 2,94	0,002*	1,21 ± 3,08

p : probabilité que les différences observées entre les moyennes adjacentes soient liées au hasard

* différence statistiquement significative

Dans le groupe participant, nous observons des différences statistiquement significatives entre les résultats du prétest et ceux du post-test (C₁), et cela, pour chacun des effets attendus du programme. Les écarts entre les résultats des tests avant et après correspondent aux effets attendus du programme.

TABEAU 19.2**Résultats descriptifs et analyses comparatives des résultats des groupes témoin et participant (n = 84)**

Variables	Groupe témoin			Groupe participant		
	Prétest $\bar{x} \pm s$	C ₂ p	Post-test $\bar{x} \pm s$	Prétest $\bar{x} \pm s$	C ₁ p	Post-test $\bar{x} \pm s$
Causes individuelles	0,278 ± 0,452	0,810	0,286 ± 0,486	0,423 ± 0,537	0,000*	0,038 ± 0,194
Causes environnementales	0,944 ± 0,787	0,008*	0,679 ± 0,716	0,712 ± 0,75	0,004*	1,15 ± 0,916
Actions individuelles	0,122 ± 0,298	1	0,183 ± 0,382	0,232 ± 0,396	0,008*	0,008 ± 0,053
Actions organisationnelles	1 ± 0,911	0,047*	0,714 ± 0,803	0,731 ± 0,992	0,000*	1,5 ± 0,828
% responsabilité employeur	6,82 ± 3,02	0,000*	7,59 ± 2,67	5,45 ± 3	0,000*	9,64 ± 1,55
% responsabilité travailleuse	4,80 ± 3,27	0,001	3,22 ± 2,94	5,71 ± 2,84	0,000*	1,21 ± 3,08

p : probabilité que les différences observées entre les moyennes adjacentes soient liées au hasard

* différence statistiquement significative

Parmi les témoins, nous observons que les moyennes avant et après (C_2) de quatre variables font l'objet de différences statistiquement significatives, ce qui pourrait indiquer la présence d'un biais d'accoutumance au test pour ces variables. Le nombre moyen de causes environnementales (avant = 0,944; après = 0,679) et d'actions organisationnelles (avant = 1; après = 0,714) diminue entre le prétest et le post-test, allant à l'inverse des objectifs visés par le programme de formation. Par ailleurs, le pourcentage de responsabilité que les témoins attribuent à l'employeur augmente (avant = 6,82; après = 7,59) alors que celui qui est attribué à la travailleuse diminue (avant = 4,8; après = 3,22), ce qui correspond à des effets attendus du programme.

Enfin, la comparaison des résultats du post-test entre les groupes participant et témoin (C_4) révèle la présence de différences significatives entre les groupes, et cela, pour chacune des six variables dépendantes. Les différences sont conformes à la théorie sous-jacente au programme.

L'exposition au programme de formation est la seule variable associée de manière statistiquement significative à la probabilité qu'un sujet attribue une cause individuelle à l'accident lors du post-test (voir tableau 19.3). Nos résultats indiquent que le programme réduit cette probabilité, le ratio de cotes étant de 0,141 et les sujets ayant participé au programme ayant une probabilité 85,9 % inférieure aux sujets du groupe témoin d'attribuer une cause individuelle à l'accident après la formation. Deux variables permettent d'expliquer la probabilité qu'un sujet mentionne une cause environnementale. Il s'agit des résultats au prétest et de l'exposition au programme. Une fois l'effet des résultats du prétest contrôlé, cette probabilité est quatre fois plus grande pour le groupe qui a participé au programme que pour le groupe témoin. En somme, il semble que le programme de formation atteigne bien l'un de ses objectifs, soit le fait d'accroître la compréhension du rôle des facteurs environnementaux dans l'occurrence des accidents de travail.

TABEAU 19.3**Résultats des analyses de régression logistique (n = 84)**

Variables dépendantes	Variables retenues	B	Erreur	p	Ratio de cotes
Causes individuelles	Exposition au programme	-1,96	0,805	0,015	0,141
<i>Valeurs observées / valeurs prédites : 84,5 %</i> <i>Rapport de vraisemblance : 64,385</i>					
Causes environnementales	Résultats prétest	1,251	0,418	0,003	3,493
	Exposition au programme	1,453	0,565	0,010	4,276
<i>Valeurs observées / valeurs prédites : 70,2 %</i> <i>Rapport de vraisemblance : 86,298</i>					
Actions organisationnelles	Résultats prétest	1,467	0,547	0,007	4,334
	Exposition au programme	2,669	0,780	0,001	14,421
<i>Valeurs observées / valeurs prédites : 78,6 %</i> <i>Rapport de vraisemblance : 66,39</i>					

Aucune variable ne permet d'expliquer les résultats obtenus au post-test concernant la probabilité de suggérer des actions individuelles qui auraient pu contribuer à prévenir cet accident. Par contre, le programme de formation multiplie par 14 la probabilité que les sujets trouvent au moins une cause organisationnelle, après que l'effet des résultats du prétest a été contrôlé.

TABEAU 19.4**Résultats des analyses de régression linéaire multiple (n = 84)**

Variables dépendantes	Variables retenues	B ajusté	p	R² ajusté	F	p
Responsabilité de l'employeur (%)	Exposition au programme	0,548	0,000	0,304	13,062	0,000
	Résultats prétest	0,34	0,001			
	Formation SST antérieure	0,221	0,033			
Responsabilité de la travailleuse (%)	Exposition au programme	-0,371	0,000	0,178	10,001	0,000
	Résultats prétest	0,31	0,003			

Trois variables sont positivement associées, et de manière statistiquement significative, au pourcentage de responsabilité attribué à l'employeur en post-test (voir tableau 19.4). Elles expliquent 30,4 % de la variation de cette variable. La participation au programme est la variable qui semble avoir l'effet le plus important, suivie des résultats au prétest, puis de la présence d'une formation. Enfin, deux variables expliquent 17,8 % de la

variation du pourcentage de responsabilité attribué à la travailleuse: la participation au programme a un effet négatif sur la variable dépendante, alors que les résultats du prétest lui sont associés de manière positive. Ces résultats semblent indiquer que le programme de formation favorise une attribution externe à l'égard des accidents, ce qui est conforme à sa théorie sous-jacente.

Discussion

Nos analyses bivariées révèlent que le groupe participant présente un plus grand nombre de changements statistiquement significatifs entre les mesures du prétest et celles du post-test. Toutes les différences vont dans le sens des objectifs de la formation. Cependant, seules les analyses multivariées permettent de vérifier si les changements observés chez le groupe participant peuvent être attribués au programme de formation.

Les résultats de ces analyses indiquent que le programme de formation étudié est associé, de manière statistiquement significative, à cinq de ses six effets attendus. Le programme entraînerait, tel qu'escompté, une amélioration de la compréhension des facteurs qui contribuent aux accidents. Cette amélioration se traduit par une moins grande probabilité, chez les sujets du groupe participant au programme, d'attribuer des causes individuelles à l'accident décrit dans notre instrument de mesure, et par une plus grande probabilité pour ces sujets de lui attribuer des causes environnementales. Le programme favoriserait également l'application des connaissances enseignées à l'égard de la prévention des accidents. Cet apprentissage se traduit par une plus grande probabilité, pour les sujets du groupe participant, de suggérer une action organisationnelle susceptible de prévenir l'accident décrit. Enfin, le programme de formation augmente le pourcentage d'attribution externe de l'accident et réduit le pourcentage d'attribution interne, conformément à l'un des objectifs poursuivis.

La présence d'une formation antérieure en SST est associée au pourcentage de responsabilité attribué à l'employeur. Cet effet attendu du programme était le seul pour lequel nous avons observé, lors des analyses bivariées, une différence significative entre les résultats du prétest des groupes participant et témoin. Or, le groupe témoin comportait une plus grande proportion de sujets formés en SST que le groupe participant. Il

était donc possible qu'un biais de sélection puisse affecter la validité interne de nos inférences concernant le pourcentage de responsabilité attribué à l'employeur. En outre, la moyenne obtenue par les témoins, en post-test, est significativement supérieure à celle du prétest. Tout se passe comme si, en réfléchissant de nouveau aux questions posées, on avait ajusté le tir. Ces résultats semblent indiquer la présence d'un biais d'accoutumance au test chez les sujets qui avaient déjà participé à une formation en SST. Cela signifierait qu'il y avait une interaction entre la sélection et l'accoutumance au test pour le pourcentage de responsabilité attribué à l'employeur. L'inclusion de la présence d'une formation antérieure en SST dans nos analyses multivariées nous a permis non seulement d'en estimer l'effet, mais aussi de faire en sorte que l'estimation de l'effet du programme de formation ne soit pas altérée par l'interaction entre la sélection et l'accoutumance au test.

Par ailleurs, nos analyses ne nous ont pas permis de vérifier la présence d'une relation entre le programme et la probabilité qu'un sujet trouve des actions individuelles susceptibles de prévenir l'accident décrit dans notre instrument de mesure. En outre, aucune variable de contrôle n'est associée à cette variable dépendante. Il est plausible que l'absence de relation statistiquement significative entre le programme et cet effet attendu soit liée à une puissance statistique insuffisante, puisque l'écart entre les résultats avant et après était faible.

Nous avons diminué la probabilité de biais d'histoire et de maturation en limitant au minimum, soit trois jours, le temps écoulé entre les mesures du prétest et celles du post-test. Le biais d'histoire a aussi été limité en raison du fait que les sujets ne retournaient pas dans leur milieu de travail entre le prétest et le post-test. Les effets de ces biais ont également été contrôlés lors de nos analyses multivariées par l'inclusion des résultats du prétest des groupes participant et témoin à titre de variables de contrôle, et par la comparaison des résultats du post-test entre ces groupes.

Nous avons tenté de réduire au minimum les biais pouvant affecter la validité externe de l'étude en recrutant des sujets dans quatre groupes. Les programmes de formation auxquels ils ont participé ont été dispensés par des formateurs différents. Il est probable que nos résultats puissent être extrapolés à l'ensemble des programmes de base en santé et en sécurité du travail dont la centrale syndicale est responsable, et cela, pour les raisons suivantes : nos taux de réponse étaient élevés (65 % et 70 %), et dans

le cadre de la phase I de notre étude, nous avons observé peu d'écart entre le programme prescrit et celui qui était effectivement implanté. Cet écart ne concernait que le temps consacré aux activités portant sur les thèmes abordés dans le cadre du présent chapitre, la durée observée des activités étant supérieure à celle qui était prescrite dans le cahier du formateur.

Malheureusement, nous ne pouvons comparer nos résultats à ceux d'autres études, puisqu'il s'agit de la seule recherche évaluative portant sur un programme de formation syndicale dont la théorie sous-jacente ait été décrite. Nos résultats ne pourraient donc être extrapolés qu'à des programmes de formation qui feraient appel à un processus similaire à celui du programme étudié. Lors de la phase I, l'analyse du verbatim de la session que nous avons enregistrée sur bande sonore nous a permis de constater que les formateurs s'opposaient aux participants lorsque ceux-ci exprimaient des opinions par lesquelles ils manifestaient leur désaccord quant aux effets attendus du programme. En général, l'interaction prenait fin lorsque le participant adoptait l'opinion du formateur. Il semble donc y avoir une concordance avec la théorie éducative à laquelle la centrale syndicale se réfère (Blondin, 1980 ; Freire, 1983).

Toutefois, le processus du programme que nous venons de décrire risque d'entraîner un biais susceptible d'affecter la validité externe de notre étude. Il s'agit du désir de plaire à l'évaluateur. En effet, comme les formateurs ont dû recourir à des méthodes de confrontation, il est probable que les participants connaissaient les effets attendus du programme et qu'ils aient été tentés de se conformer aux objectifs de la formation lorsqu'ils ont rempli le questionnaire du post-test. Cependant, le fait que les réponses aux questionnaires aient été anonymes pourrait limiter la portée de cet effet. Seules des observations sur les actions syndicales menées par les participants à leur retour dans leur milieu de travail permettraient de vérifier la valeur prédictive des mesures à court terme des effets de la formation que nous avons effectuées.

Lorsque nous avons présenté nos résultats à la centrale, nous avons souligné avec insistance la plausibilité de la théorie sous-jacente au programme et l'efficacité du programme. Nous avons recommandé la poursuite de la formation sans modification, si ce n'est de réévaluer le temps prescrit pour les activités relatives au thème des accidents de travail.

Le climat de confiance qui s'est développé avec les permanents syndicaux au cours des années de collaboration ayant précédé cette recherche

a grandement facilité le processus de l'évaluation. Nous n'avons jamais senti de résistance de la part du personnel de la centrale à l'endroit de nos travaux. Au contraire, les membres ont manifesté un intérêt réel et nous ont facilité la tâche lorsque nous avions besoin de leur collaboration. Notre seul regret est de ne pas avoir obtenu les ressources financières suffisantes pour terminer cette recherche rapidement. Les organismes subventionnaires font face à un paradoxe : leurs dirigeants souhaitent que les résultats de recherche soient utilisés par les milieux de pratique, et pour favoriser une telle utilisation, les chercheurs doivent travailler en partenariat avec des représentants de ces milieux. Or, ces derniers sont généralement peu enclins à participer à nos recherches si nous ne nous engageons pas à produire des connaissances à courte échéance. L'absence de ressources financières adéquates peut compromettre le respect d'un tel engagement.

Pour aller plus loin :

CHEN, H. T. (1990), *Theory-driven Evaluations*, Newbury Park, CA, Sage Publications.

Ce livre a été rédigé à la suite du ralentissement abrupt aux États-Unis des évaluations de programmes sociaux, dans les années 1980. Cette période faisait suite à deux décennies au cours desquelles les évaluations concluaient fréquemment que les programmes n'atteignaient pas leurs objectifs. L'approche de la boîte noire s'étant révélée peu utile, Chen insiste sur l'importance de posséder des connaissances sur l'objet du programme à évaluer avant de tenter d'en analyser les effets. Il innove en intégrant, dans un même cadre théorique, les théories normative et causale des programmes. Une lecture essentielle pour les évaluateurs qui souhaitent acquérir des habiletés pour conceptualiser les théories des programmes.

CHEN, H. T. (2005), *Practical Program Evaluation. Assessing and Improving Planning, Implementation, and Effectiveness*, Thousand Oaks, Sage Publications.

Il s'agit du plus récent ouvrage de Chen. Il y propose une série d'outils conceptuels et opérationnels pour évaluer et répondre aux besoins des groupes d'intérêt concernés par l'évaluation de programmes. Son contenu est centré sur l'action, du développement de la théorie du programme à l'évaluation de ses effets. Plutôt que de proposer des recettes, l'auteur énonce les principes sur lesquels les stratégies d'évaluation doivent s'appuyer et fonde ses recommandations sur des bases théoriques explicites.

SHADISH, W. R., T. D. COOK et D. T. CAMPBELL (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, Houghton Mifflin Company.

Ce livre s'inscrit dans l'école de pensée développée par Donald T. Campbell, qui a publié avec J. C. Stanley *Experimental and Quasi-experimental Designs for Research* (1963), puis, avec Thomas D. Cook, *Quasi-experimentation: Design and Analysis Issues for Field Settings* (1979), chez Rand-McNally. C'est un ouvrage de référence incontournable pour ceux et celles qui souhaitent connaître les méthodes qui permettent d'optimiser la validité interne et externe des inférences de causalité. Les auteurs présentent les biais susceptibles d'altérer la validité, selon les méthodes utilisées, et proposent des solutions pour éviter qu'ils n'apparaissent ou pour en vérifier la présence.

Internet

Institut de recherche Robert-Sauvé en santé et en sécurité du travail.

Le site suivant permet de télécharger les rapports de recherche des études financées par l'IRSST; quelques recherches évaluatives sont disponibles: www.irsst.qc.ca

20

Une évaluation des effets d'un programme de réduction des conflits chez les enfants

Françoise Fortin et François Bowen

Le contexte

En juin 2001, à la suite d'un appel d'offres, le Centre national de prévention du crime du Canada (CNPC, ministère de la Justice) a mandaté notre groupe de recherche, le Groupe d'étude sur la médiation en milieu scolaire (GEMMS), pour effectuer l'évaluation de la deuxième édition d'un programme de médiation par les pairs et de résolution de conflits portant le nom de *Vers le pacifique*. L'appel d'offres stipulait que la mise en œuvre, les effets ainsi que le rapport coût-bénéfice du programme devaient être évalués sur une période de trois années à l'aide d'un devis de type expérimental. L'évaluation demandée portait spécifiquement sur l'efficacité du programme. L'organisme qui financerait l'évaluation avait également contribué financièrement au développement de l'intervention, et souhaitait connaître le plus précisément possible les effets du programme tel qu'appliqué rigoureusement par le personnel des milieux scolaires (Flay *et al.*, 2005). De plus, le programme devait être évalué dans plusieurs types de milieux (urbain, rural, semi-rural) afin de pouvoir le disséminer à travers le Canada si les résultats se révélaient satisfaisants. Tout comme

la validité interne de l'étude, contrôlée grâce au devis expérimental, la question de la validité écologique des résultats, soit la possibilité de transférer les résultats obtenus à d'autres contextes, était importante pour les bailleurs de fonds.

Bien que la première édition de *Vers le pacifique* eût fait l'objet d'une évaluation (1995-1997), il était aussi nécessaire d'évaluer cette deuxième édition, car des modifications importantes avaient été apportées au programme. En effet, le nombre d'ateliers proposés, la clientèle ciblée (ajout d'un programme visant les élèves de maternelle), l'organisation et la mise en place du programme (sur deux années plutôt que sur une seule) avaient été modifiés. De plus, l'évaluation précédente ne s'était pas suffisamment attardée à l'étude des effets proximaux du programme (par exemple, les connaissances acquises sur la résolution de conflit), pas plus qu'elle n'avait investigué sur ses retombées dans l'environnement scolaire (par exemple, les climats relationnel et de sécurité). Enfin, le premier devis ne couvrait que deux années.

Les principaux acteurs

Chaque organisme subventionnaire a ses exigences quant aux acteurs à impliquer dans une évaluation et à l'importance de leur participation. De son côté, lorsqu'il attribue une subvention, le CNPC demande la mise en place d'un comité consultatif constitué de représentants des principaux partis impliqués. Pour l'évaluation de *Vers le pacifique*, ce comité était constitué, outre du chercheur principal, d'un co-chercheur, de la coordonnatrice de recherche, d'un chargé de dossier du CNPC, du directeur de la recherche au Centre international de résolution de conflits et de médiation (CIRCM) / Institut Pacifique (les concepteurs du programme), d'un représentant des commissions scolaires dans lesquelles le programme était mis en œuvre ainsi que d'un professeur-chercheur indépendant. Ce comité se réunissait environ deux fois par année afin de discuter du devis et des rapports préliminaires d'évaluation de la mise en œuvre et des effets du programme. Toute modification majeure au devis ou au plan d'analyse devait être approuvée par le comité.

Par ailleurs, une entente entre les chercheurs du GEMMS et le CIRCM/Institut Pacifique a été conclue pour favoriser le recrutement des écoles, mais aussi le partage de l'information, particulièrement en ce qui a

trait à l'évaluation de la mise en œuvre. Les concepteurs du programme étaient donc assez engagés dans le processus d'évaluation. Enfin, des communications régulières avaient lieu entre le groupe de recherche et les différents milieux scolaires afin de discuter des modalités d'évaluation ou de répondre à des questions sur le déroulement des activités de recherche.

Le cadre logique du programme

Le but du programme *Vers le pacifique* est de prévenir la violence par la promotion de conduites pacifiques. Il s'adresse aux enfants de la maternelle à la 6^e année du primaire. Étendue sur deux années, l'implantation du programme se scinde en deux volets successifs, soit la résolution de conflits et la médiation par les pairs.

Le premier volet, les ateliers de résolution de conflits, vise à développer chez les jeunes diverses habiletés leur permettant d'établir des relations interpersonnelles pacifiques. La formation se fait au moyen d'ateliers ayant comme objectifs de favoriser chez les jeunes une vision non violente des situations conflictuelles, de les encourager à acquérir des habiletés d'écoute, de jugement critique, d'expression verbale des émotions et de maîtrise de soi, de développer une meilleure compréhension ainsi qu'une meilleure gestion de leurs conflits, et d'aider les jeunes à améliorer leurs relations avec autrui et leur compréhension d'eux-mêmes et des autres. L'ensemble de ces objectifs est abordé par l'entremise de 9 ateliers pour chacun des cycles (20 à la maternelle) conçus pour être animés par les enseignants titulaires.

Une fois les ateliers mis en place, l'école peut s'engager dans le deuxième volet du programme, la médiation par les pairs. Les jeunes y apprennent à utiliser la médiation comme mode de résolution de conflits. Cela nécessite la sélection, la formation et l'encadrement d'élèves médiateurs. Les objectifs de ce volet sont de reconnaître que les jeunes ont le pouvoir de résoudre des conflits et de développer leurs compétences dans ce sens, d'amener les jeunes à utiliser la médiation comme alternative efficace aux approches infructueuses de résolution de conflits, d'encourager les jeunes à s'impliquer dans la résolution de conflits, et de les responsabiliser dans la gestion de leurs propres conflits. Enfin, parallèlement à l'évaluation, les concepteurs du programme ont développé un troisième volet appelé Famille-Communauté, qui ne sera pas abordé ici.

TABLEAU 20.1

**Modélisation des résultats attendus :
schéma des principaux éléments évalués**

Description du programme	Éléments de la mise en œuvre	Impacts proximaux et médiaux (attendus)	Impacts distaux (attendus)
Volet Ateliers de résolution de conflits (1 ^{re} année d'implantation) • 20 ateliers à la maternelle • 9 ateliers pour chaque année du primaire	Première année A) Formation et soutien des enseignants B) Préparation d'ateliers C) Arrimage des contenus avec les autres interventions réalisées à l'école D) Arrimage du programme avec le contexte organisationnel de l'école E) Comités de coordination	Première et deuxième années (éventuellement, troisième année pour le maintien des acquis) (par ordre d'acquisition) A) Connaissances et habiletés accrues en lien avec la résolution de conflits B) Accroissement des conduites prosociales C) Diminution des conduites agressives (objectif médial) D) Diminution des problèmes internalisés (objectif médial) E) Réduction des taux généraux d'agressivité et de victimisation (objectif médial)	Deuxième et troisième années A) Amélioration du climat relationnel et de sécurité entre les élèves B) Capacité d'appliquer en situation réelle, seul ou avec l'aide d'un médiateur, le processus de résolution pacifique de conflits C) Amélioration du climat organisationnel D) Sentiment accru de compétence chez les enseignants
Volet Médiation (2 ^e année d'implantation) • Ateliers de rappel des contenus de la première année • Sélection, formation et encadrement des médiateurs	Deuxième et troisième années F) Formation et soutien des enseignants G) Sélection, formation et encadrement des médiateurs H) Action des médiateurs + points B, C, D et E		

Le tableau 20.1 illustre la modélisation des effets attendus du programme. Dans un premier temps, le modèle suppose l'acquisition ou l'accroissement de connaissances et d'habiletés en résolution de conflits (1^{re} année), nécessaires au développement des conduites prosociales et de résolution de conflits (1^{re} et 2^e années). Le développement des conduites prosociales et de résolution de conflits comme solutions de rechange à la violence devrait progressivement mener à une diminution des conduites agressives (2^e année) et, conséquemment, à réduire les taux généraux d'agressions et de victimisation à l'échelle de l'école (vers la 3^e année). Enfin, cette diminution des taux d'agressions devrait favoriser la perception d'un climat scolaire plus positif au point de vue de la sécurité et des relations sociales (2^e et 3^e années).

Le protocole d'évaluation

Les objectifs visés et les questions d'évaluation

Tel que mentionné précédemment, l'appel d'offres du CNPC impliquait une évaluation de la mise en œuvre et des effets du programme, mais aussi une analyse coût-bénéfice. Pour les besoins de la démonstration, nous ne nous attarderons ici que sur l'évaluation des effets.

La principale, et ultime, question concernant l'évaluation des effets était la suivante : Est-ce que le programme *Vers le pacifique* est efficace pour réduire la fréquence des comportements antisociaux et la violence chez les jeunes ? Plus précisément, il s'agissait d'évaluer les retombées des ateliers et de la médiation par les pairs sous plusieurs angles, soit les impacts proximaux (par exemple, l'acquisition de connaissances et d'habiletés en résolution de conflits), médiaux (par exemple, les conduites agressives) et distaux (par exemple, le climat relationnel).

Le devis

Un protocole de recherche longitudinale ayant été retenu, la collecte des données s'est déroulée sur quatre années scolaires. Il a par ailleurs été convenu qu'un devis quasi expérimental serait employé. Celui-ci devait permettre de comparer les effets des programmes de huit écoles expérimentales et de cinq écoles témoins. Le tableau 20.2 présente le devis d'évaluation sous forme schématique. Les écoles ont été choisies en fonction de leur milieu – urbain, rural ou semi-rural – afin d'optimiser les possibilités de généralisation des résultats de l'évaluation. Comme les écoles n'ont pas toutes été réparties au hasard entre les groupes expérimental et témoin, un devis quasi expérimental a dû être substitué au devis expérimental véritable qui devait être employé.

TABLEAU 20.2**Schématisation du devis de recherche**

	1 ^{re} année	2 ^e année	3 ^e année	4 ^e année
Expérimental	X O X	X O ₂ X	X O ₂ X	O ₂ X
Témoin	X X	X X	X O X	O ₂ X

X: prise de mesure

O: ateliers

O₂: ateliers et médiation

Plusieurs instruments, sources et méthodes de collecte de données ont dû être utilisés pour arriver à évaluer les divers objectifs du programme. Le tableau 20.3 détaille ces outils.

TABLEAU 20.3**Instruments utilisés**

Dimensions évaluées	Instruments	Auteurs	Sources
Autocontrôle comportemental Conduites agressives Conduites prosociales Habilités en résolution de conflits Problèmes internalisés (retrait et isolement social)	Échelles sociométriques et comportementales (questionnaires)	Tremblay <i>et al.</i> (1992) Bowen <i>et al.</i> (2003)	Enseignants Parents Pairs
Connaissances relatives aux attitudes et aux comportements requis pour résoudre pacifiquement des conflits	Questionnaire sur les connaissances	Bowen <i>et al.</i> (2003)	Auto-rapporté, tous les élèves et les membres du personnel
	Entrevue de résolution de problèmes interpersonnels (ERPI)	Groleau (1990)	Auto-rapporté, 5 élèves par classe
Climat relationnel Climat de sécurité Victimisation perçue Victimisation agie Victimisation vécue	Questionnaire sur l'environnement socioéducatif (QES)	Janosz <i>et al.</i> (2007)	Auto-rapporté, tous les élèves

Le plan de valorisation et l'utilisation potentielle des résultats

Les résultats de l'évaluation devaient servir à mesurer les effets attendus du programme en lien avec le processus de mise en œuvre, afin de déterminer quelles étaient les conditions qui amélioraient l'efficacité des interventions promues et celles qui leur faisaient obstacle. À terme, le commanditaire de cette évaluation souhaitait qu'elle lui permette de déterminer la pertinence

de disséminer le programme dans tout le pays. Par ailleurs, plusieurs mécanismes de transfert et de valorisation des résultats ont aussi été mis en place. Premièrement, après chaque année d'évaluation, ainsi qu'à la toute fin du processus, un rapport était déposé auprès du CNPC et discuté en comité consultatif. Deuxièmement, chaque direction d'école du groupe expérimental recevait annuellement un rapport d'évaluation de la mise en œuvre du programme dans son milieu, accompagné d'un certain nombre de recommandations ou de suggestions visant à accroître la qualité de son implantation. Troisièmement, toutes les écoles qui en faisaient la demande pouvaient recevoir annuellement un bilan décrivant l'environnement socioéducatif de leur école (climats, pratiques...). Quatrièmement, les chercheurs du GEMMS ont diffusé les résultats de l'évaluation dans des congrès professionnels et scientifiques tout au long des quatre années du projet. Enfin, comme les directions d'école, les concepteurs (CIRCM / Institut Pacifique) du programme recevaient les divers rapports d'évaluation, en plus d'être en contact régulier avec le groupe de recherche, notamment pour assurer une mise en œuvre optimale du programme.

Le budget et la planification

Le budget initial de l'évaluation était de 557 000 \$ répartis sur trois années d'évaluation. La plus grande part de cette somme a servi à payer les salaires des assistants de recherche et de la coordonnatrice pour la collecte et la saisie des données nécessaires à l'évaluation des effets et de la mise en œuvre.

Le processus de l'évaluation

Le contexte entourant le recrutement des écoles

Le recrutement des écoles participantes constituait un aspect particulièrement important du processus d'évaluation, car la représentativité géographique des sujets (zones urbaines, banlieues et zones rurales) et les possibilités de généralisation des résultats seraient tributaires de la qualité de l'échantillonnage.

Au cours du mois d'août 2001, l'équipe du GEMMS a obtenu de la part du CNPC une réponse positive concernant la soumission déposée à la

mi-juillet, à la suite de l'appel d'offres. Le financement demandé avait été octroyé, mais le projet devait débiter rapidement, soit en septembre de la même année. Concrètement, le recrutement des écoles était fait conjointement avec les membres du CIRCM / Institut Pacifique, qui présentaient le programme auprès des commissions scolaires et des établissements d'enseignement. Cette présentation s'accompagnait d'une description du projet de recherche et de ses exigences. Respecter le plan initial de recrutement et d'échantillonnage, qui se voulait au départ aléatoire stratifié, aurait demandé plusieurs semaines, voire des mois. Or, le temps nous manquait cruellement, compte tenu de l'appel d'offres tardif (juin) pour un projet d'une telle envergure, et qui devait de surcroît se déployer dès septembre.

Le devis stipulait que les 16 écoles sélectionnées devaient provenir de trois types de milieu, soit urbain, rural et semi-rural. Après quelques hésitations, il a été convenu de recruter les écoles dans trois commissions scolaires de l'île de Montréal et de sa couronne nord. Nous avons rencontré la direction et le personnel de toutes les écoles sélectionnées, essentiellement entre les mois d'août et de décembre 2001, pour leur faire part des implications (intervention et évaluation) liées à la participation au projet mais aussi des avantages inhérents à cette participation. Le projet a été présenté à toutes les écoles suivant les mêmes procédures: l'école devait accepter de s'engager sans savoir si elle bénéficierait de l'intervention dès les deux premières années ou seulement au cours des deux années subséquentes. À la suite de la présentation, les représentants des écoles ont signé une lettre d'engagement envers le CIRCM / Institut Pacifique. Cependant, à cause des négociations des enseignants du secteur public sur le dossier de l'équité salariale alors en cours, une mesure de boycott des activités parascolaires a été appliquée dans certaines des écoles recrutées. Ce phénomène, combiné avec la résistance de quelques établissements, a provoqué le retrait de trois écoles témoins. Ainsi, des 16 écoles initialement prévues, seules 13 ont finalement accepté de faire partie de l'échantillon. Parmi elles, huit expérimentaient le programme depuis deux ans (les deux premières années du programme), et les cinq autres étaient des écoles témoins. En théorie, nous aurions pu poursuivre notre recrutement afin d'atteindre nos objectifs de randomisation de l'échantillon, mais deux obstacles nous en ont empêchés. Non seulement le temps nous manquait, mais il aurait aussi pu être contestable, d'un point de vue éthique, de n'offrir à ces nouvelles écoles que la possibilité de faire partie

du groupe témoin, contrairement aux écoles du recrutement initial, qui avaient eu le choix. De plus, le poids des moyens de pression déployés dans les écoles durant cette période, ainsi que certaines exigences des commissions scolaires, ne nous permettaient pas d'espérer de meilleurs résultats même en poursuivant la procédure de recrutement.

La répartition des sujets

Le devis expérimental choisi initialement prévoyait la répartition aléatoire des sujets (écoles) entre les groupes expérimental et témoin. Ce choix méthodologique aurait permis de rendre les groupes équivalents, réduisant ainsi les menaces à la validité interne de l'étude et renforçant la qualité des inférences causales. En effet, en contrôlant les variables potentiellement confondantes (par exemple, le niveau de défavorisation), les effets obtenus sont théoriquement attribuables uniquement au programme ou à sa mise en œuvre. Quoique très pertinent sur papier, ce type de devis amène souvent des difficultés lors de son application dans la réalité. Ce problème s'est justement posé dans notre évaluation. En effet, certaines commissions scolaires ont fait des pressions afin de choisir, à titre d'écoles expérimentales, les écoles qu'elles considéraient comme étant celles qui avaient le plus besoin du programme (problèmes de violence importants, milieux très défavorisés). Étant donné les contraintes temporelles évoquées plus haut, l'équipe d'évaluation n'a pas eu d'autre choix que d'accepter ces conditions. L'utilisation d'un devis quasi expérimental sans répartition aléatoire des sujets s'est donc imposée. Cela a cependant eu des implications importantes pour la suite du projet, car les groupes expérimental et témoin n'étaient pas équivalents concernant certaines mesures telles que le statut socioéconomique ou l'origine ethnique des élèves.

Le choix des instruments de mesure

Dans un devis de type expérimental ou quasi expérimental, les instruments privilégiés sont validés, c'est-à-dire que leurs propriétés psychométriques ont été établies lors d'études antérieures. Le principal avantage des instruments validés est qu'ils permettent de comparer plusieurs études utilisant les mêmes instruments. Il peut cependant s'avérer difficile de trouver un instrument déjà validé qui réponde parfaitement à

l'évaluation de certains des objectifs (ou impacts) de l'intervention à examiner.

Dans le cas qui nous occupe, plusieurs des échelles de mesure provenaient d'instruments validés lors d'évaluations de programmes de même nature (par exemple, la première édition du programme *Vers le pacifique*). Parmi ces instruments, les échelles sociométriques et comportementales (Tremblay *et al.*, 1992) et l'entrevue de résolution de conflits interpersonnels (ERPI, Groleau, 1990), qui ont pu être reprises telles quelles ou légèrement adaptées. Ces instruments ou procédures ont permis d'obtenir des mesures directement associées aux impacts (proximaux, médiaux et distaux) attendus du programme tels que présentés dans le tableau 17.1.

Pour évaluer les connaissances et habiletés en résolution de conflits, l'entrevue de résolution de problèmes interpersonnels (ERPI) a été choisie, car aucun questionnaire autorévéle n'était disponible pour couvrir tous les élèves de la maternelle à la sixième année. Cependant, compte tenu des coûts élevés d'une telle procédure (passation individuelle, codage et transformation des données), nous avons sous-échantillonné en sélectionnant, sur une base aléatoire, en moyenne cinq élèves par classe. Toutefois, selon le modèle logique du programme (voir le tableau 20.1), cette mesure permet de déterminer l'atteinte d'un impact proximal important, soit la connaissance des attitudes et des comportements relatifs à la résolution de conflits. Il aurait donc été souhaitable qu'une telle évaluation puisse être réalisée pour l'ensemble de l'échantillon. Afin de résoudre ce problème (et de réduire les coûts associés à l'ERPI), le groupe de recherche a mis au point un questionnaire autorévéle dit « des connaissances », qui a été rempli par tous les élèves de la 3^e à la 6^e année. Ce questionnaire a été validé au cours des deux premières années d'implantation du programme en comparant ses résultats avec ceux de l'ERPI, et en analysant ses propriétés psychométriques (voir Bowen *et al.*, 2003 pour plus de détails). De cette manière, nous avons donc pu obtenir des mesures concernant les connaissances et habiletés en résolution de conflits acquises par l'ensemble des élèves des deuxième et troisième cycles.

La collecte de données et la prise de mesures

Dans un protocole de nature expérimentale, les données doivent être recueillies au même moment et dans des conditions semblables afin de garantir la comparabilité des résultats obtenus. Pour l'évaluation de *Vers le pacifique*, les diverses collectes des données étaient prévues à la rentrée, durant le mois d'octobre et à la fin de l'année scolaire, durant le mois de mai, puisque le programme devait être implanté essentiellement entre les mois de novembre et d'avril. Il s'est avéré difficile de respecter scrupuleusement cet échéancier, et cela pour plusieurs raisons.

Tout d'abord, à cause des délais très serrés pour effectuer le recrutement des écoles, la collecte initiale (T₁) des données a dû être étalée de novembre 2001 à janvier 2002. Et puis, le fait de devoir obtenir le consentement des parents pour la participation de leurs enfants au projet ne faisait qu'accroître ce délai. En effet, les enfants étant mineurs, le comité d'éthique de l'université ainsi que ceux des différentes commissions scolaires ont demandé aux parents de signer un formulaire de consentement pour que leurs enfants participent à l'évaluation du programme. Il avait cependant été convenu avec les représentants des écoles et du CIRCM/Institut Pacifique que les interventions auprès des élèves ne débuteraient qu'après la collecte des données.

Ensuite, les écoles n'avaient que des disponibilités limitées pour permettre la collecte des données. En outre, il était impossible de récolter toutes les données en une seule journée, chaque école nécessitant entre deux et quatre journées de travail. Enfin, il n'a pas toujours été facile de trouver des auxiliaires de recherche durant la collecte de l'automne, puisque la plupart d'entre eux étaient des étudiants universitaires qui suivaient des cours pendant cette période.

Pour toutes ces raisons, le groupe de recherche a choisi de répartir le travail des auxiliaires sur deux mois (octobre et novembre) plutôt que sur un seul, comme cela avait été prévu initialement. Cet étalement de la collecte a pu faire en sorte que, dans certains cas, la collecte des données a été réalisée alors que l'intervention venait à peine de débuter. Il est cependant difficile d'estimer dans quelle mesure ce léger accroc au protocole a pu biaiser les résultats des analyses d'impacts.

Par ailleurs, le GEMMS a tenté de mettre en place un certain nombre de procédures afin de rendre les conditions de passation les plus semblables

possible d'une classe à l'autre. Pour les passations de groupe effectuées par les auxiliaires de recherche (désignations comportementales, questionnaire sur les connaissances), les questions étaient lues à voix haute aux élèves, et ce, à plusieurs reprises. De plus, le calme et le silence devaient régner dans les classes. Pour ce qui est de la passation du questionnaire sur l'environnement socioéducatif (QES), réalisée par les enseignants titulaires, ceux-ci devaient faire la collecte dans une autre classe que la leur afin de réduire le plus possible la désirabilité sociale chez les élèves, car ces derniers devaient répondre à des questions concernant le climat dans leur classe et les pratiques de leur enseignant. Les enseignants devaient également lire les instructions qui décrivaient très précisément la procédure à suivre.

La passation de l'ERPI était pour sa part réalisée par l'auxiliaire de recherche, qui rencontrait un à un les élèves ayant été sous-échantillonnés pour les fins de l'entrevue. Afin de diminuer les biais possibles, les auxiliaires chargés de la collecte des données avaient reçu une formation à cet effet et bénéficiaient en tout temps de la présence d'un coordonnateur prêt à offrir son aide et son soutien en cas de besoin. Enfin, des consignes détaillées étaient données aux enseignants et aux parents pour que ceux-ci puissent remplir leurs questionnaires respectifs.

En dépit du fait que les procédures de collecte des données ont été appliquées avec une grande rigueur, certains problèmes n'ont pu être résolus. En premier lieu, comme le projet de recherche s'est déroulé sur une période de quatre années, il s'est avéré difficile d'assurer la pérennité de l'équipe de recherche sur une si longue période. Il a donc fallu former régulièrement de nouveaux auxiliaires de recherche ainsi que des coordonnateurs. De plus, il fallait tenir compte du fait que les auxiliaires engagés, des étudiants, provenaient de différents horizons disciplinaires, comme le droit, la psychoéducation, les études internationales, et qu'un très grand nombre d'entre eux faisaient des études de premier cycle. La formation et l'encadrement de ces auxiliaires devaient donc prendre en compte le peu d'expérience du milieu scolaire de certains et le peu d'expertise méthodologique des autres.

En second lieu, certaines situations particulières causées par des contingences plus ou moins contrôlables ont pu affecter la qualité de certaines données recueillies, bien qu'il soit hasardeux d'estimer l'ampleur de ces biais. Ce fut par exemple le cas de la passation du QES, pour laquelle

un petit nombre d'enseignants n'avaient pas lu les directives et n'avaient donc pas changé de classe avec un collègue, ou encore certaines entrevues (ERPI) se sont déroulées dans des endroits achalandés et parfois bruyants (faute d'espace adéquat disponible).

Finalement, un facteur de nature contextuelle (ou historique) a influencé la collecte des données durant les deux premières années du projet, soit la grève du zèle de plusieurs enseignants menée dans le cadre des moyens de pression exercés au sujet du dossier de l'équité salariale. Un des moyens de pression utilisés était la non-participation des enseignants à toute activité qui n'était pas explicitement inscrite dans leurs tâches, par exemple, la réponse aux divers questionnaires requis pour l'évaluation du programme. Craignant qu'une telle situation ne menace le recrutement des écoles, le chercheur principal a demandé l'appui du syndicat des enseignants (la CEQ), un des partenaires majeurs du centre de recherche dont faisait partie ce chercheur, afin qu'une directive vienne appuyer le projet en raison de ses très nombreuses retombées positives. De plus, il a proposé aux écoles du groupe témoin, outre l'implantation du programme après deux ans, la possibilité de bénéficier d'autres activités ou services qui ne visaient pas les mêmes objectifs que le programme. Malgré cela, cette situation a eu comme conséquence la diminution du nombre d'écoles témoins recrutées de huit à six, ainsi que le désistement d'une autre école témoin après la première collecte de données. Dans les écoles qui avaient accepté de participer à l'évaluation, certains enseignants ont refusé à la dernière minute de remplir les questionnaires ou ne les ont pas remplis en entier, ou encore ont refusé de faire la passation du QES auprès des élèves. Malgré ces problèmes non négligeables qui ont pu affecter la robustesse des résultats sur le plan statistique, il n'en demeure pas moins que la lettre écrite par la haute direction de la CEQ a sans doute contribué à sauver le projet.

Les principaux résultats de l'évaluation

Afin de faciliter la lecture, la section suivante est séparée en deux parties en fonction du modèle logique du programme, qui distingue les effets proximaux (les gains attendus chez les enfants au cours des deux premières années, concernant les connaissances relatives à la résolution de conflits et la mise en pratique des comportements qui y sont associés) des

effets médiaux et distaux (la réduction attendue des problèmes externalisés, comme l'agressivité, et internalisés, comme le retrait social, au cours des deux dernières années de mise en œuvre du programme). Le lecteur qui souhaite obtenir des résultats plus détaillés et davantage d'information sur les procédures statistiques utilisées pourra consulter le rapport final d'évaluation du programme (Bowen *et al.*, 2006).

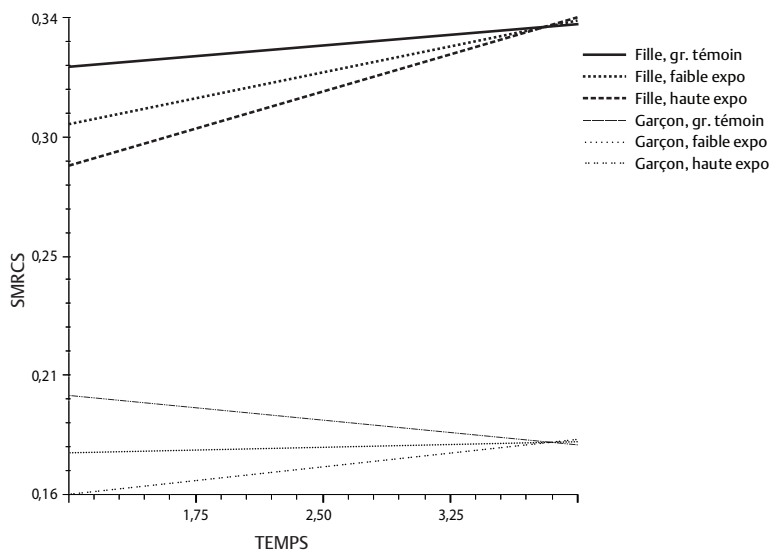
Les impacts proximaux

Après deux années de mise en œuvre, les élèves des écoles expérimentales présentent des gains significatifs en ce qui a trait à la gestion des émotions lors de situations conflictuelles ($\beta = 0,44$; $t = 9,408$; $p < 0$), à la capacité d'identifier les causes d'un conflit ($\beta = 0,057$; $t = 6,84$; $p < 0$), aux solutions de réparation proposées en vue de résoudre un problème ($\beta = 0,049$; $t = 3,083$; $p < 0,05$), à l'expression des émotions ($\beta = 0,0033$; $t = 4,27$; $p < 0$), aux habiletés de résolution de conflits ($\beta = 0,0044$; $t = 4,8$; $p < 0$), à l'autocontrôle comportemental ($\beta = 0,0028$; $t = 3,398$; $p < 0,0001$) et aux conduites prosociales ($\beta = 0,002$; $t = 2,557$; $p < 0,005$).

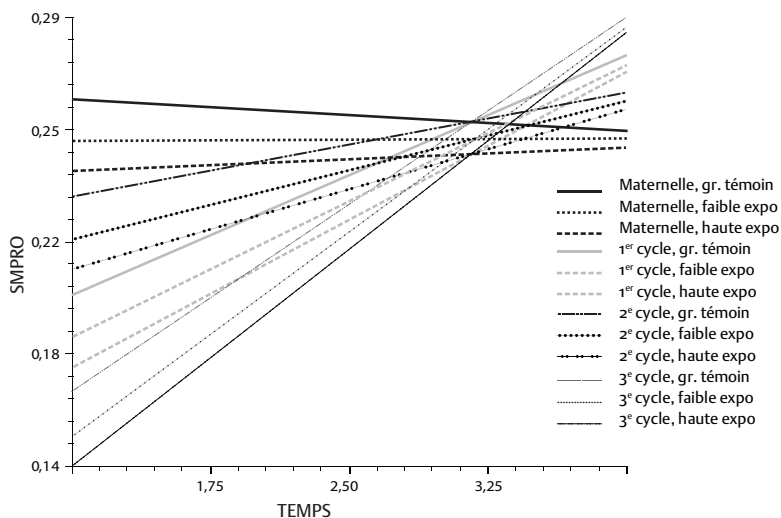
Les gains sont plus importants chez les filles que chez les garçons, et cela pour toutes ces variables, à l'exception des solutions de réparation proposées en vue de résoudre un conflit. Les élèves plus jeunes (maternelle et premier cycle) bénéficient davantage du programme pour ce qui est de la gestion des émotions, des habiletés de résolution de conflits (voir figure 20.1), des conduites prosociales et de l'autocontrôle comportemental. Seule la capacité d'identifier les causes d'un conflit est meilleure chez les élèves des 2^e et 3^e cycles. Quant aux autres variables, aucune différence en fonction de l'âge n'est apparue. Enfin, les élèves ayant un niveau élevé d'exposition au programme ont obtenu des résultats plus concluants, en ce qui a trait aux conduites prosociales, que les autres élèves (voir figure 20.2).

FIGURE 20.1

Effet différentiel du programme sur les habiletés de résolution de conflits (SMRCS) en fonction du sexe et du niveau d'exposition

**FIGURE 20.2**

Effet différentiel du programme sur les conduites prosociales (SMPRO) en fonction de l'âge et du niveau d'exposition



Les impacts médiaux et distaux

Après trois années de mise en œuvre du programme, on constate que les garçons plus jeunes ayant un niveau élevé d'exposition au programme présentent significativement moins de problèmes externalisés que les élèves du groupe de comparaison. Les résultats ne sont cependant pas significatifs pour les filles et les élèves plus âgés des deux sexes. On constate en revanche une hausse de ces problèmes chez les élèves ayant un degré d'exposition peu élevé au programme. Par ailleurs, les résultats montrent une réduction des problèmes internalisés chez les élèves ayant un niveau élevé d'exposition au programme. Tout comme pour les problèmes externalisés, les problèmes internalisés se sont avérés en hausse chez les élèves ayant un niveau d'exposition peu élevé à l'intervention.

À la lumière de ces résultats, on constate que les hypothèses du modèle logique du programme sont en partie confirmées. En effet, le changement dans les conduites agressives semble être proportionnel au développement de connaissances et d'habiletés de résolution de conflit (identification des causes d'un conflit, solutions de réparation proposées en vue de résoudre un conflit). Dans un deuxième temps, ces connaissances sont effectivement utilisées et appliquées par les élèves (autocontrôle comportemental, gestion et expression des émotions, habiletés de résolution de conflits), ce qui semble avoir pour effet de favoriser leurs conduites prosociales. Enfin, la mise en place de solutions de rechange à la violence (habiletés en résolution de conflits et prosociales) entraînerait une diminution des conduites agressives chez les élèves.

Les leçons apprises : forces et limites de l'évaluation

Premièrement, dès le début du projet, notre groupe de recherche a mis en place un partenariat avec les concepteurs du programme. Cette alliance a sans nul doute constitué l'une des forces importantes de notre évaluation. Sans ce partenariat, il aurait été impossible de recruter à temps toutes les écoles nécessaires pour l'évaluation du programme. De plus, l'implication du CIRCM / Institut Pacifique a grandement facilité la collecte des données, son personnel travaillant de concert avec le groupe de recherche afin que les questionnaires destinés aux enseignants soient remplis le plus rapidement possible. Les concepteurs du programme ont également participé au développement du questionnaire sur les connaissances transmises, permet-

tant ainsi de mieux cibler les gains attendus de l'intervention. Ce partenariat s'est aussi avéré bénéfique pour le CIRCM / Institut Pacifique, qui a pu profiter de l'expertise d'une équipe de chercheurs universitaires pour produire une évaluation rigoureuse de son programme. L'évaluation de la mise en œuvre et des effets de *Vers le pacifique* a permis de formuler une série de commentaires et de recommandations concernant les éléments efficaces du programme et ceux qui exigeaient des ajustements. Comme tout partenariat, celui-ci a exigé des engagements importants de part et d'autre, mais les bénéfices mutuellement obtenus en valaient bien la peine.

Deuxièmement, en ce qui concerne le modèle d'évaluation choisi, une des grandes limites des modèles expérimentaux est leur manque de flexibilité, tant sur le plan du traitement proposé que sur celui de la sélection des sujets ou des instruments. Dans le cadre de notre projet, ce modèle ayant posé certaines contraintes, le groupe de recherche a dû apporter des modifications au devis initial. Ce fut le cas, par exemple, à propos de la répartition aléatoire des sujets, qui n'était pas réaliste dans le contexte de cette évaluation (plusieurs écoles se seraient alors désistées, manifestant clairement le désir de faire partie du groupe expérimental). L'élaboration du questionnaire sur les connaissances est un autre exemple du manque de souplesse d'un tel devis. Compte tenu de l'évolution des apprentissages, certaines questions posées aux élèves auraient dû être ajustées au fur et à mesure que les acquisitions se faisaient, car un « effet plafond » a été constaté pour certaines échelles de mesure de ce questionnaire après deux ou trois ans. Cependant, si de telles modifications avaient été effectuées, les mesures n'auraient plus été comparables, d'un point de vue statistique.

L'approche expérimentale choisie avait néanmoins plusieurs forces qui se doivent d'être mises en évidence. Tout d'abord, le fait d'utiliser un devis de cette nature permet de faire des attributions causales beaucoup plus facilement, à savoir que les effets observés sont dus au programme et non à d'autres variables confondantes telles que le développement naturel des compétences chez les enfants. Les résultats concernant les retombées de l'intervention apparaissent ainsi suffisamment fondés pour que le CNPC recommande la dissémination du programme à travers les écoles du Canada. De plus, les mesures utilisées, provenant d'instruments validés, facilitent la comparaison avec d'autres programmes de même nature. Cela permet au CNPC, par exemple, de subventionner le programme ayant les effets les plus probants sur le plan de la prévention de la violence.

Par ailleurs, quoique peu fréquente dans les devis expérimentaux, une évaluation de la mise en œuvre particulièrement minutieuse a été effectuée afin de connaître le véritable traitement auquel les enfants avaient été exposés, car la qualité de la mise en œuvre était assez variable selon les écoles du groupe expérimental, sans compter le fait que les écoles témoins pouvaient offrir à leurs élèves d'autres activités visant des objectifs se rapprochant de ceux de *Vers le pacifique*. Sans une évaluation aussi approfondie de la mise en œuvre du programme, accompagnée d'une investigation sur les interventions menées dans les écoles du groupe témoin, il aurait été très difficile d'interpréter les résultats des analyses statistiques.

Tout comme pour l'évaluation de la mise en œuvre du programme, nous avons compris l'importance de documenter le processus de l'évaluation même, afin de connaître les variables confondantes possibles, telles que des conditions de passation des questionnaires différentes ou l'impact des rétroactions sur la qualité de la mise en œuvre du programme données aux milieux par notre équipe. Cette précaution a été très utile, surtout sur le plan de l'analyse des données, pour comprendre pourquoi certaines données manquaient (refus de participer ou hasard) ou encore certains résultats non attendus (par exemple, l'augmentation des conduites agressives). Nous avons donc réussi à avoir une certaine flexibilité dans l'application du modèle expérimental, mais cela tout en documentant les biais que cette souplesse pouvait induire, précaution qui nous a permis de conserver notre rigueur méthodologique.

En ce qui concerne le devis utilisé par notre groupe de recherche, il se basait sur plusieurs sources (parents, pairs, enseignants, enfants) et comportait deux méthodes principales de collecte des données. L'accessibilité à plusieurs sources et méthodes a donc permis de réduire le taux d'erreurs de mesures associé à chacune des sources prises individuellement (Shadish *et al.*, 2002), en plus de connaître le point de vue de différents acteurs sur les acquisitions des élèves. Les résultats ont en effet montré que, pour certaines mesures, seuls les pairs rapportaient un changement, alors que pour d'autres mesures, seul l'enseignant observait une différence. Enfin, étant donné le rapport coût-bénéfice avantageux des questionnaires, l'évaluation a pu se faire sur un grand nombre de sujets, ce qui augmentait le potentiel de généralisation des résultats et permettait de détecter des effets plus subtils (meilleure puissance statistique). Tous ces éléments seront à conserver en cas d'évaluation ultérieure.

Somme toute, le modèle expérimental utilisé, avec les modifications qui lui ont été apportées, a permis de répondre aux demandes du CNPC, qui désirait pouvoir porter un jugement sur l'efficacité de ce programme dans les conditions d'une application rigoureuse par les milieux scolaires. Ce modèle s'avère particulièrement efficace lorsqu'on veut établir des attributions causales entre un programme et les effets observés, et lorsque ce programme et la modélisation de ses effets attendus sont bien développés, comme c'était le cas dans le cadre de l'évaluation du programme *Vers le pacifique*.

Une évaluation d'implantation au Burkina Faso avec une méthode mixte

*Valéry Ridde, Seni Kouanda, Maurice Yameogo,
Kadidiatou Kadio et Aristide Bado*

L'UNICEF (2009) avance que 80 % des décès maternels pourraient être évités si les interventions connues étaient mises en œuvre. Ainsi, dans le domaine de la lutte contre la mortalité maternelle, de nombreuses voix s'élèvent pour déplorer que des stratégies *a priori* efficaces ne soient pas mises en place, notamment en Afrique (De Brouwere et Van Lerberghe, 2001 ; United Nations Secretary-General, 2010). L'une des interventions recommandées est de favoriser le recours à du personnel qualifié lors des accouchements. Si les soins à l'accouchement étaient bien dispensés, les risques de mortalité et de morbidité maternelles seraient réduits de 95 % (Graham, Bell et Bullough, 2001 ; World Health Organization, 2010a). Or, les recherches ont montré que la barrière financière était l'une des multiples causes du faible recours des femmes aux accouchements assistés par du personnel qualifié (Nanda, 2002). Ainsi, il est maintenant recommandé de réduire, voire de supprimer, le paiement direct des accouchements (African Union, 2010 ; Richard, Witter et de Brouwere, 2010 ; World Health Organization, 2010b).

Cependant, il ne suffit pas de suivre les recommandations internationales en formulant des politiques publiques, encore faut-il que ces dernières soient effectivement mises en œuvre, car « if implementation fails, everything fails » (Chen, 2004). Or, d'autres auteurs affirment que « many evidence-based innovations fail to produce results when transferred to community in the global south, largely because their implementation is untested, unsuitable or incomplete » (Madon, Hofman, Kupfer et Glass, 2007). En outre, les méta-analyses montrent que l'efficacité des interventions est généralement soumise à de multiples facteurs de distorsion liés à la mise en œuvre, dépassant souvent les 50 % (Durlak et DuPre, 2008). Malheureusement, les recherches sur la mise en œuvre des politiques publiques sont rares en Afrique, seulement 4 % des publications entre 1933 et 2003 (Saetren, 2005), et peu s'intéressent aux enjeux d'équité en santé (Gilson et Raphaely, 2008). De plus, peu de travaux en promotion de la santé ou sur les politiques de santé dans les pays à faibles revenus font directement référence aux concepts et théories utilisés pour l'étude des politiques publiques (Breton et De Leeuw, 2011 ; Gilson et Raphaely, 2008).

Contexte et intervention

Quarante six pour cent de la population du Burkina Faso est considérée comme vivant sous le seuil de pauvreté, avec un revenu de 0,41 euros par jour. On estime que 4 000 femmes meurent chaque année à la suite de troubles liés à leur maternité (World Health Organization [WHO], United Nations Children's Fund [UNICEF], United Nations Population Fund [UNFPA] et World Bank, 2010) et les interventions *a priori* efficaces pour réduire la mortalité maternelle sont encore peu couramment effectuées (World Health Organization, 2010a). Ainsi, seulement 43 % des femmes accouchaient dans une formation sanitaire publique en 2006 (Direction des études et de la planification, 2007). Pour s'attaquer à ce problème, le Burkina Faso a décidé d'appliquer en 2007 une politique publique de subventions du prix des accouchements. La politique est entièrement financée par l'État (Ministère de la Santé, 2006b). Depuis janvier 2007, le coût d'un accouchement eutocique évalué par le Ministère à 4 500 F CFA (6,86 euros) est subventionné à 80 % dans les maternités des centres de santé de première ligne. Les femmes ne doivent plus payer que les 20 % restants, soit 900 F CFA (1,4 euros). Ce montant est considéré comme un forfait prenant en compte

toutes les dépenses médicales liées à l'accouchement pratiqué dans une maternité. Le manuel d'exécution de la politique précise que ce forfait concerne les « actes, médicaments et consommables, mise en observation » (Ministère de la Santé, 2006a). Il recommande le contenu des intrants du « kit accouchement » et précise le protocole clinique que le personnel de santé doit suivre en cas d'accouchement.

Plusieurs études qualitatives ont montré que la formulation de cette politique n'avait pas été suffisamment inclusive. Elles révèlent également que de nombreuses femmes ainsi que le personnel de santé connaissent mal cette politique. Dans ces études, quelques entrevues qualitatives indiquent aussi que les femmes paient plus que le forfait officiel de 900 F CFA (Amnesty International, 2009 ; Ridde *et al.*, 2011 ; Sombié *et al.*, 2007). Certaines recherches ont déjà étudié la mise en œuvre de la politique, mais cependant aucune étude populationnelle n'a encore mesuré de manière systématique cet écart de paiement et, le cas échéant, tenté de l'expliquer.

Ainsi, les objectifs de cette étude sont 1) de mesurer les dépenses effectuées par les femmes lors d'un accouchement eutocique dans une maternité de première ligne ; et 2) d'expliquer les résultats au moyen d'une enquête qualitative. Notre étude porte donc spécifiquement sur le déficit de mise en œuvre du paiement du forfait pour les accouchements et non pas sur la mise en œuvre de la politique en général.

Méthodes

L'approche méthodologique est celle d'une évaluation d'implantation descriptive et analytique assortie d'un devis mixte séquentiel (Creswell et Plano Clark, 2007). La description (puis la discussion) de la méthode tient compte des conseils visant à accroître la qualité de la description du recours aux méthodes mixtes (Pluye *et al.*, 2011).

Lieu de l'étude. L'étude se déroule dans le district sanitaire de Ouaragaye, majoritairement rural, de 260 000 habitants vivant d'une agriculture de subsistance. Cinquante-cinq pour cent des habitants de ce district du Centre-Est vivaient sous le seuil de pauvreté en 2003. La société et le système de santé sont organisés de manière très pyramidale. La tendance au maintien du *statu quo* l'emporte sur l'instauration de réformes, notamment en faveur de l'équité. Le personnel de santé exerce de l'ascendant

sur les populations, et la participation communautaire se limite le plus souvent à la participation financière (Ridde, 2008).

Évaluation de la mise en œuvre. Il existe de multiples approches et peu de consensus sur les concepts susceptibles de servir à analyser les écarts entre ce qui est prévu et ce qui est réalisé (Durlak et DuPre, 2008 ; Patton, 2008 ; Saetren, 2005 ; Walt *et al.*, 2008), c'est-à-dire sur l'*implementation gap* (Pressman et Wildavsky, 1984), le déficit de mise en œuvre. Notre étude évalue non pas des processus en vue de comprendre la dynamique du jeu des acteurs dans la mise en œuvre, mais plutôt l'implantation, et elle est centrée sur la mesure et sur l'analyse des écarts. Elle porte sur la politique publique de subvention des accouchements. Contrairement à l'approche utilisant des cadres d'analyse souvent dictés par le concept et les caractéristiques de la fidélité (Carroll *et al.*, 2007), notre étude repose sur une démarche explicative fondée sur quatre pistes de recherche que nous décrivons plus loin.

Méthodes mixtes. Un devis mixte séquentiel explicatif a été employé (voir chapitre 7) en donnant une importance égale aux données qualitatives et quantitatives (Creswell et Plano Clark, 2007). L'approche a été séquentielle, car l'utilisation de méthodes quantitatives a précédé et éclairé le recours subséquent à des données qualitatives. Le devis a été explicatif : les données qualitatives collectées ont servi à expliquer et à comprendre les résultats obtenus avec les données quantitatives. Après l'analyse des données quantitatives, les chercheurs ont défini quatre modes d'explication de l'écart constaté entre le forfait officiel et le paiement déclaré par les femmes. Ils ont tenu compte de contexte, de la politique mise en œuvre et de la littérature scientifique sur le sujet : 1) le paiement de produits non prévus dans le « kit accouchement » donnant droit au forfait officiel ; 2) le paiement de produits non offerts à la pharmacie du centre de santé ; 3) la difficulté éprouvée par les responsables de la mise en œuvre à comprendre la politique ; 4) les comportements répréhensibles des agents de santé.

Le volet quantitatif. Une étude transversale avec une enquête ménage a été réalisée en février 2010. La sélection des femmes ayant accouché dans des maternités a été réalisée à partir des registres des maternités. Les enquêteurs ont recherché les noms de toutes les femmes ayant accouché dans les 25 formations sanitaires ($n = 1016$) au cours des six dernières semaines. Grâce à des informateurs clés dans les villages, 926 femmes ont

été trouvées. Cinquante autres femmes ayant accouché dans un CSPS mais dont les noms ne figuraient dans aucun registre ont été trouvées ainsi que 71 parturientes ayant accouché à domicile. Au total, 1035 femmes ont été interrogées. Le questionnaire provenait d'une précédente recherche et avait été adapté. Il portait sur les caractéristiques démographiques et socio-économiques des ménages et les dépenses médicales engagées lors de l'accouchement. Une double saisie des données a été effectuée avec EpiData. Les analyses descriptives ont été réalisées avec les logiciels SPSS et Excel.

Le volet qualitatif. Pour examiner de manière critique les quatre modes d'explication des résultats de l'enquête quantitative, une enquête qualitative a été conduite entre décembre 2010 et janvier 2011. Les données qualitatives ont été intégrées aux données quantitatives au moment de l'interprétation des résultats. Dans un premier temps, un *focus group* a réuni 33 personnes (dont une seule femme), soit tous les responsables des maternités ou des centres de santé, plus trois membres de l'équipe cadre du district sanitaire. Les chercheurs ont présenté les résultats quantitatifs, et on a mené ensuite une discussion sur les quatre modes d'explication. Les noms des centres de santé n'étaient pas donnés. La discussion a permis aux chercheurs d'accéder à un premier niveau d'analyse réflexive. Puis, pour compléter et enrichir les données obtenues et pour tenir compte des biais inhérents aux discussions de groupe (contrôle social), spécialement en Afrique (Olivier de Sardan, 2008), nous avons réalisé des entrevues individuelles en profondeur. De façon à mieux comprendre le phénomène (voir chapitre 10), elles se sont déroulées dans des maternités considérées comme des cas extrêmes. Suivant le principe des devis séquentiels dans lesquels les données quantitatives déterminent le choix des méthodes qualitatives (Creswell et Plano Clark, 2007), les cas en question ont été sélectionnés à partir des résultats de l'enquête quantitative. Ainsi, des entrevues ont été effectuées dans quatre des cinq maternités (trois rurales et une urbaine) où la moyenne du paiement déclaré par les femmes était la plus grande (> 2 300 F CFA). La cinquième maternité n'a pas été visitée car le responsable avait été muté depuis l'enquête quantitative. Des entrevues ont été effectuées avec les principaux responsables de la mise en œuvre (*implementor*) de la politique. Nous appuyant sur notre connaissance du contexte et sur des études conduites antérieurement en Afrique (Agyepong et Nagai, 2011; Olivier de Sardan, 2009; Walker et Gilson, 2004), nous avons jugé que seuls les agents de première ligne dans les

centres de santé (*street level bureaucrats*) pouvaient être parfaitement renseignés sur la politique. En effet, les femmes pouvaient nous dire *combien* elles payaient (ce qu'elles ont fait lors de l'étude quantitative), mais elles ne pouvaient pas expliquer *pourquoi* elles payaient. Ainsi, nous avons rencontré quatre infirmiers responsables des centres de santé, quatre responsables des maternités, deux agents itinérant de santé, une accoucheuse villageoise et quatre responsables de pharmacie communautaire. De plus, deux entrevues individuelles ont été effectuées avec des responsables du district, l'une avec le responsable actuel de la santé de la reproduction qui coordonne les interventions dans ce domaine, et l'autre, avec l'ancien médecin chef du district qui était présent au moment de l'enquête quantitative. Toutes ces entrevues visaient à recueillir des données concernant les quatre modes d'explication. Le *focus group* et les entrevues ont été enregistrés puis transcrits à l'aide d'un logiciel de traitement de texte. L'analyse des données a été réalisée de manière manuelle selon une approche inductive et au moyen de regroupements thématiques (Miles et Huberman, 1994).

Résultats

Nous considérerons d'abord les résultats quantitatifs et nous nous attacherons ensuite à étudier les résultats qualitatifs.

La moitié des femmes déclarent payer plus que le forfait officiel

La description de l'échantillon quantitatif est présentée dans le tableau 21.1.

Le tableau 21.2 présente les dépenses déclarées par les femmes pour chacune des rubriques concernées. Ainsi, trois ans après l'instauration de la politique nationale, 50 % des femmes déclaraient payer plus que le forfait officiel de 900 F CFA normalement réclamé pour un accouchement eutocique. La moyenne du paiement déclaré était deux fois plus élevée que le forfait et représentait sept jours de revenu pour 46 % de la population vivant sous le seuil de pauvreté. Dans aucune maternité du district, la moyenne des dépenses médicales n'était inférieure au forfait. Cette moyenne variait de 1 075 F CFA (1,64 euros) au minimum à un maximum de 4 232 F CFA (6,45 euros) (figure 21.1). Nous n'avons pas cherché à comprendre pour le moment cette hétérogénéité. Trente pour cent des femmes

interrogées déclaraient avoir acheté des médicaments à la pharmacie communautaire, et 6 % auprès du personnel de santé.

TABEAU 21.1

Caractéristiques des femmes de l'enquête quantitative en 2010

Âge moyen	26,14
Statut matrimonial	
Célibataire	20 (1,99 %)
Mariée monogame	528 (52,49 %)
Mariée polygame	416 (41,35 %)
Union libre	38 (3,78 %)
Veuve/divorcée	4 (0,40 %)
Éducation	
Aucune	950 (94,43 %)
Primaire	38 (3,78 %)
Secondaire et plus	18 (1,79 %)
Distance des CSPS par rapport au domicile	
< 5 km	564 (55,7 %)
5 < 10 km	277 (27,3 %)
> 10 km	172 (17,0 %)

TABEAU 21.2

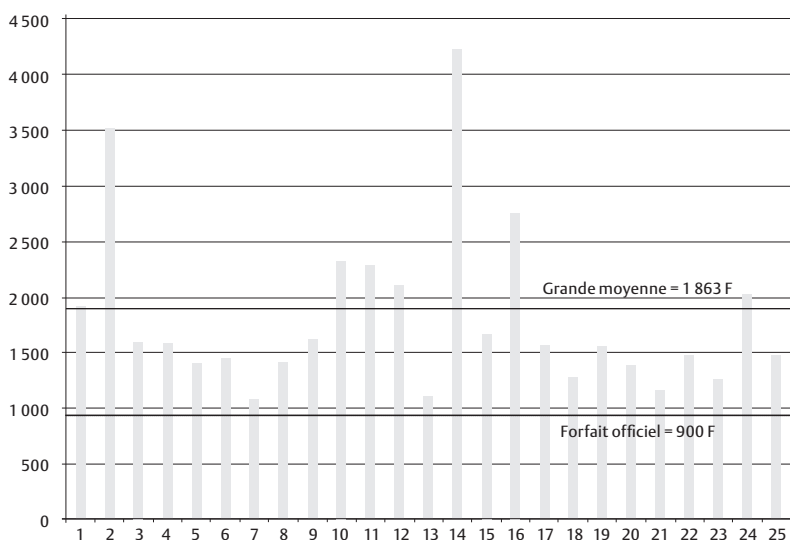
Dépenses médicales (F CFA) déclarées pour les accouchements eutociques dans le district de Ouargaye (2010)

	N*	Médiane	Moyenne	Écart-type	Minimum	Maximum
Ticket d'accouchement	N = 883	0	0	0	0	0
Frais d'accouchement	N = 883	700	819	290	0	17 400
Frais d'hospitalisation	N = 883	200	108	122	0	1 400
Frais de médicaments soignant	53	1 100	1 786	1 552	200	6 600
Frais de médicaments dépôt	266	1 500	2 917	4 399	100	50 000
Total	883	900	1 863	2 814	0	50 900

* La colonne N correspond au nombre de personnes ayant réalisé ces dépenses pour chaque items, expliquant qu'il soit différent à chaque fois.

FIGURE 21.1

**Dépenses moyennes en F CFA selon les 25 maternités
du district de Ouargaye (2010)**



Comprendre l'écart constaté

En plus d'une remise en cause des résultats quantitatifs par le personnel de santé, quatre autres facteurs peuvent expliquer le déficit lié à la mise en œuvre du forfait car, dit un agent de santé, « *ce n'est pas normal* ».

Nier l'écart et remettre en cause l'enquête

Tous les agents de santé ont manifesté spontanément et immédiatement de la surprise au moment de la présentation des résultats quantitatifs (« Ça nous étonne beaucoup, ce chiffre-là, franchement », a dit un infirmier), mais surtout ils ont douté de la véracité des conclusions. Les chercheurs ont fait face à une contestation des résultats : « Vous êtes certains de vos enquêtes ? » a dit un infirmier.

D'abord, on a douté de la capacité des femmes à répondre aux questions concernant les dépenses : « Vraiment cet écart-là, bon, ça m'étonne. Je me demande si réellement les femmes, le jour de l'interrogatoire, elles ont compris », avance un infirmier. D'autres étaient d'avis que les femmes

interrogées avaient mal compris les questions et confondu certaines choses, comme par exemple les contours du forfait.

« Il y a le cas des faux travaux. La femme vient par exemple et pense que c'est pour un accouchement, on décèle d'autres pathologies. Et après les soins, la femme accouche. Elle va penser que [la politique] prend en charge tous ces trucs-là », dit un infirmier.

Certains agents ont accusé les femmes d'avoir déclaré de façon mensongère avoir payé plus que nécessaire parce qu'elles pensaient qu'elles seraient remboursées : « Et puis nos parents, là, ils pensent que, quand elles vont dire ça là, vous allez les rembourser », affirme une accoucheuse.

Ensuite, c'est la démarche méthodologique qui a été remise en cause, et en particulier la validité des données ou l'objectivité des réponses :

« Avez-vous exigé de voir le reçu ? Est-ce que réellement les personnes interrogées au moment de l'enquête, est-ce qu'elles n'ont pas dit ce qu'elles voulaient dire seulement ? » demande un infirmier.

Bien que les analyses quantitatives aient concerné uniquement les femmes ayant eu un accouchement eutocique, certains agents se sont demandé si les chercheurs n'avaient pas inclus dans les résultats des dépenses déclarées celles des accouchements dystociques pour lesquels le forfait dans ce type d'accouchement est quatre fois plus élevé (voir plus loin).

« Kit ouvert » ou « kit fermé » ?

Pour bien comprendre la situation au moment de l'enquête qualitative (mais aussi les difficultés liées à la mise en œuvre de la politique), le personnel de santé a voulu nous expliquer en détail la gestion des intrants pour les accouchements. Au moment de l'instauration de la politique en 2007, les responsables du district ont choisi d'utiliser un fonctionnement qu'ils qualifiaient de « kit ouvert ». Les agents ne sont pas obligés de sortir du stock tous les produits officiellement inscrits dans le manuel pour le « kit accouchement » (Ministère de la Santé, 2006a), ils doivent noter seulement ceux qu'ils utilisent réellement. Puisque le Ministère rembourse les établissements sanitaires pour la valeur du forfait (3 600 F CFA) en plus du paiement effectué par les femmes (900 F CFA), moins on utilise de produits et plus on fait des économies. Un infirmier se souvient : « On

a constaté qu'y avaient des femmes qui pouvaient accoucher avec une seule paire de gants qui coûtait 50 F CFA. » Un système de primes avait été institué dans ce district où la différence monétaire (en moyenne 1 000 F CFA = 1,52 euros) entre le forfait remboursé par le Ministère et les dépenses réelles était répartie à 80 % pour le comité de gestion communautaire du centre de santé et 20 % pour les agents de santé. Tous les groupes et individus (femmes, agents, comité) visés par la politique étaient « *contentes* », pour reprendre le mot du responsable qui avait organisé cette stratégie. Ce mode de fonctionnement a duré un an. Puis, l'organe central du Ministère a exigé l'application de la politique nationale, c'est-à-dire le retour au « kit fermé » et l'arrêt des primes. Mais deux ans après, ce même organe central s'est rendu compte que des fonds étaient gaspillés du fait d'un remboursement forfaitaire dépassant les dépenses réelles et a prescrit le « kit ouvert ». Les centres de santé sont donc depuis lors remboursés en fonction de leurs dépenses réelles. Mais au moment de l'enquête quantitative, c'était la formule du « kit fermé » qui était en place.

Le paiement de produits non inclus dans le kit officiel

Dans toutes les maternités, en dehors de la lessive pour laver la salle d'accouchement que les femmes doivent payer, le personnel de santé confirme qu'il a prescrit des produits nécessaires aux parturientes mais ne faisant pas partie du kit donnant droit au paiement du forfait. On a donné des médicaments contre le paludisme et des antibiotiques aux femmes ou encore du collyre aux nouveau-nés, tel que cette citation l'illustre bien :

« Pour moi, s'il y a paludisme c'est en dehors de [la politique], je te prescris. Si je vois que ça nécessite une voie veineuse, je te prescris une voie veineuse. Franchement le coût va changer », précise un agent de santé.

Mais la compréhension n'est pas la même pour tous (voir plus loin), car un infirmier nous a affirmé au contraire que les médicaments contre le paludisme pendant l'accouchement étaient compris dans la politique.

Dans une seule maternité (urbaine), les agents ont eu aussi tendance à affirmer que ce paiement supplémentaire s'expliquait par des prescriptions réclamées par les femmes. Par exemple : « Il y a certaines femmes qui demandent qu'on leur achète des produits pour les aider à accoucher facilement », dit une accoucheuse.

Il faut enfin ajouter que, dans un centre de santé, on facture systématiquement aux femmes un montant supplémentaire de 300 F CFA (0,46 euros) pour l'acte de déclaration de naissance de l'enfant. Dans les autres centres, les parents doivent se rendre dans les mairies des communes rurales pour effectuer ce paiement, s'ils le souhaitent.

Le paiement de produits non offerts à la pharmacie du centre de santé

Il semble que les ruptures de stock dans les pharmacies des centres de santé soient très rarement à l'origine des dépenses supplémentaires effectuées par les femmes. Le contexte spécifique et conjoncturel d'une seule maternité rurale évoque cette situation. On rappelle l'absence régulière du gérant de la pharmacie communautaire et la présence d'une pharmacie privée dans le village pour justifier la possibilité de ces paiements supplémentaires : « Si le gérant n'est pas là, il faut aller au dépôt privé pour cela », dit l'agent de santé.

La difficulté à comprendre les mesures politiques

« Nous, on a un problème sérieux de compréhension de [la politique] », nous a dit un infirmier. Trois exemples illustreront le manque d'information.

D'abord, dans toutes les maternités, les accouchements eutociques qui nécessitent une épisiotomie étaient considérés par le personnel de santé comme des accouchements dystociques. Or, ces derniers obligent les femmes à payer un montant forfaitaire de 3600 F CFA et non pas 900 F CFA. Ainsi, un infirmier se souvient « qu'au début l'épisiotomie était considérée comme dystocique et que la femme payait alors 3600 F CFA ».

Ensuite, dans toutes les maternités, le moment où le forfait doit être appliqué a des répercussions sur l'achat de produits supplémentaires. Par exemple, le forfait ne serait applicable qu'à partir du moment où le travail de la femme commence et il s'arrête après la naissance de l'enfant, sauf pour les hémorragies prises en charge dans les sept jours. Ainsi, tous les intrants utilisés avant ou après ce temps sont facturables à la parturiente, et il en est de même des gants d'examen (servant à vérifier si le travail a commencé), de l'antispasmodique ou des médicaments destinés à soigner un épisode fébrile de l'accouchée qui est toujours à la maternité.

Enfin, cette difficulté de compréhension s'est étendue à la composition des « kits d'accouchement » : « Les gens n'avaient pas les mêmes compréhension des kits », se souvient un infirmier. De plus, les agents ne savaient pas quoi faire des produits formant le kit qu'ils n'ont pas utilisés lors de l'accouchement.

Trois raisons principales sont invoquées pour expliquer cette difficulté de compréhension. La première concerne l'accompagnement offert par les responsables de la politique : ou bien ils ne soutiennent pas les *implementors* (« Au niveau national, on n'a pas eu de directives qui disaient que c'est dystocique ou eutocique », dit un médecin), ou bien leurs conseils sont trop changeants (« Ce n'est pas stable, tantôt ils disent que c'est comme ci, tantôt ils disent que c'est comme ça », dit une accoucheuse en milieu rural). La deuxième raison est liée à l'information. De la formation a bien été donnée au début de l'application de la politique, mais aucune formation continue ni aucun document d'information n'a été prévu. Ainsi, les agents nouvellement recrutés, telle cette accoucheuse dans son premier poste, ne sont pas au courant : « Je n'ai aucune idée concernant cela, comme moi je suis nouvelle là, mais ça fait dix mois. » Un autre agent en place depuis longtemps constate : « On n'a pas encore eu de documents qui nous orientent. » La troisième raison serait liée à une stratégie intentionnelle, que nous évoquerons plus loin. Un responsable affirme : « C'est une faille que les agents utilisent quand, par exemple, la femme dit qu'elle a payé l'amoxiciline par-ci par-là, les agents disent que non, que ce n'était pas lié à sa grossesse. »

Des mauvais comportements des agents de santé

Il est évidemment difficile de disposer de données empiriques explicites sur les mauvais comportements des agents de santé bien qu'un médecin dise : « Si les femmes payent plus de 900 F, ce sont des comportements liés à l'individu, ce n'est pas officiel. » Cependant, l'enquête quantitative a révélé que seulement 56 % des femmes disposaient d'un reçu justifiant leur paiement effectué au centre de santé, ce qui était une preuve évidente.

Le discours normatif est souvent présent, comme nous l'avons vu dans le *focus group*. Mais il est possible, à partir des discours implicites, d'inférer l'existence de comportements indiquant un refus d'effectuer un paiement forfaitaire.

Cette situation est notamment vraie pour ce qui touche l'utilisation des produits disponibles dans les « kits fermés » mais non utilisés pour les accouchements. Que deviennent-ils ? « Tu fais ton accouchement [...] le restant du kit, ça c'est un problème », précise un agent de santé. Cette question est d'autant plus brûlante que l'arrêt du paiement des primes aux agents de santé et le retour au « kit fermé » peuvent évidemment avoir incité les acteurs à trouver des moyens de combler le manque à gagner. Lorsque les responsables centraux ont dit aux infirmiers au cours d'une réunion que la prime de 20 % ne devait plus être attribuée et que l'on retournait au « kit fermé », les infirmiers n'ont rien dit. Leur acceptation s'explique certainement par le fait qu'ils prévoyaient utiliser le reste des kits. Le responsable du district avait prévenu : « L'utilisation de façon perverse du kit, on ne pourra pas la contrôler. »

Les agents de santé se posent aussi des questions : « Nous venons de découvrir ces chiffres-là. On veut mieux comprendre pour revoir et situer plus ces responsabilités-là. Aussi, individuellement, on veut attraper les gens. » Un médecin ajoute : « Le problème c'est qu'il n'y a pas de surveillance et, tant que ça restera comme ça, ils vont toujours piller la population. » De plus, les agissements du responsable du district de l'époque pour empêcher de tels phénomènes de se produire montrent que l'hypothèse de revente des intrants n'est pas farfelue. Il avait en effet d'abord calculé les dépenses réelles effectuées dans les maternités pour les accouchements. C'est ainsi que, vu l'écart avec le forfait remboursé par l'État, il s'était rendu compte du gaspillage des ressources publiques. Il a changé le fonctionnement des kits. Un infirmier se souvient que ce responsable « voulait voir plus clair dans gestion du reste des kits, et [qu'] il a été demandé de passer du “kit fermé” au “kit ouvert” ». Ce responsable devait bien avoir quelques soupçons, car « quand je leur ai dit qu'il y avait une équipe qui venait pour interroger les femmes sur ce qu'elles ont payé, il y en a eu qui ont tremblé ».

Discussion

Sur la méthode

La force de la méthode quantitative est d'avoir pu confirmer et mesurer avec une enquête populationnelle l'écart de paiement que des études qua-

litatives avaient abordé (Amnesty International, 2009 ; Ridde *et al.*, 2011) et que deux recherches quantitatives (mais moins exhaustives) avaient précisé. La validité externe des résultats quantitatifs est appréciable, car deux recherches conduites dans deux districts éloignés de Ouargaye ont donné des résultats analogues pour la moyenne des paiements d'un accouchement eutocique, soit 1 400 F CFA à Nouna (De Allegri *et al.*, 2011) et 1 300 F à Djibo (Ben Ameur *et al.*, 2010). Les faiblesses de la méthode quantitative, qui est moins utile pour comprendre que pour mesurer, ont été largement compensées par l'approche qualitative subséquente. Cela montre la pertinence du recours à un devis séquentiel pour atteindre l'objectif défini dans ce chapitre. Certes, l'approche qualitative ne permet pas d'établir de liens de causalité quantifiés, mais les données recueillies et la triangulation des sources d'information ont permis de comprendre une grande partie des facteurs associés au déficit de mise en œuvre.

Discours officiels et normes pratiques

Les réactions spontanées, défensives et normatives des agents de santé à l'égard du déficit de mise en œuvre du forfait ne sont pas surprenantes. Elles tiennent d'abord à la première méthode utilisée (*focus group*) dans laquelle le contrôle social ainsi que la présence de leur supérieur hiérarchique direct ne facilitent pas l'expression relative à l'absence de respect du contenu d'une politique. Mais ces réactions normatives sont également courantes dans l'administration publique et le système de santé de la région où les logiques de connivence et d'évitement dans le traitement des problèmes ainsi que l'écart entre les normes professionnelles officielles et pratiques ont déjà été mis au jour (Booth, 2011 ; Olivier de Sardan, 2010 ; Ridde, 2011). Une étude sur la mise en œuvre d'audit de la mortalité maternelle dans les établissements sanitaires a révélé que le personnel de santé avait de la difficulté à s'autocritiquer. De plus, 40 % de ce personnel a affirmé que les femmes mentaient aux chercheurs (Richard *et al.*, 2009). Il serait donc utile de documenter en profondeur la culture professionnelle des accoucheuses pour savoir si les normes pratiques communes – éloignées des normes professionnelles et de l'interprétation culturaliste du fonctionnement de l'État en Afrique (Olivier de Sardan, 2010) – n'expliqueraient pas en partie les stratégies d'adaptation des agents de santé et l'écart constaté dans la mise en œuvre.

Une mise en œuvre perfectible

La présente recherche, effectuée dans un seul district, confirme, mesure et explique le déficit de mise en œuvre concernant le paiement du forfait qui n'est pas respecté. Puisque cette situation a aussi été observée dans d'autres districts du pays, on peut penser que le problème est national. Le rôle des ruptures de stock n'a pas vraiment été confirmé, mais les trois autres facteurs aident ensemble à la compréhension du phénomène, sans que l'on puisse désigner un facteur plutôt qu'un autre.

Contrairement à d'autres pays qui ont subventionné les soins (Meessen *et al.*, 2009 ; Ridde et Morestin, 2011), la formulation de la politique a été relativement participative et on s'est gardé de toute orientation populiste (Booth, 2011). Cependant, la participation s'est souvent limitée aux responsables centraux et périphériques du système de santé et aux gynécologues, laissant peu de place aux agents de première ligne, comme en Afrique du Sud (Walker et Gilson, 2004). Le contenu des « kits d'accouchement » ainsi que le montant du remboursement forfaitaire n'ont pas été suffisamment adaptés aux réalités locales de la première ligne. De plus, contrairement aux conseils donnés concernant la mise en œuvre de politiques publiques (Brinkerhoff et Crosby, 2002), les décideurs du Burkina Faso refusent souvent de créer des projets pilotes qui permettraient pourtant de repérer les erreurs avant de mettre la politique en application. Les phénomènes qui ont entouré les accouchements en 2007 se sont reproduits en 2010 avec le programme national de lutte contre le paludisme.

En Afrique de l'Ouest, bien souvent les responsables centraux des ministères de la Santé ne restent que peu de temps en poste. Ils sont souvent recrutés par des organisations qui peuvent leur offrir de meilleures conditions de travail, ce qui ajoute à la faible continuité administrative avant et après le départ. C'est le cas de cette politique où les hauts responsables ont très souvent changé. Entre 2005, époque où la politique a commencé à être pensée, et 2010, trois femmes se sont succédé à la tête du service responsable de cette politique. Or, l'expérience de l'abolition du paiement au Ghana (Witter et Adjei, 2007 ; Witter, Adjei, Armar-Klemesu et Graham, 2009) et de l'établissement de la couverture universelle (Mathauer et Carrin, 2011) montre l'importance de l'*institutional ownership* et du soutien de la mise en œuvre par les instances dirigeantes. Or, dans le cas de cette politique, le constat est plutôt celui du « *policy-*

driven institutional incoherence» (Booth, 2011). En 2011, le Ministère a changé son organigramme, et sa politique est définie par une nouvelle direction, ce qui aura certainement des effets sur les institutions. De plus, nous avons vu que les messages véhiculés par ces responsables, lorsqu'ils se rendaient sur le terrain, n'étaient pas toujours les mêmes, qu'ils changeaient souvent suivant les personnes. La cellule qui supervise cette politique reste trop petite, ne dispose pas d'un nombre suffisant d'experts et a de la peine à prendre en compte les données probantes pour adapter la politique, comme l'a montré le calcul du forfait des accouchements. Ainsi, les incidences du passage d'un remboursement au forfait à un remboursement des dépenses d'accouchement au réel restent à étudier. C'est un sujet sur lequel on manque cruellement de connaissances en Afrique (Ridde et Morestin, 2011).

Trop peu d'efforts continus d'information sur le contenu de la politique ont été déployés, tant à l'égard des populations que du personnel de santé. Les nouveaux agents ne sont pas au courant des détails, et le niveau d'information du personnel de santé en général est inégal, ainsi que l'a montré, par exemple, la lutte contre le VIH au Burkina Faso (Ouattara, Gruénais, Zongo et Ouedraogo, 2011). Les maternités ne disposent pas d'un document accessible expliquant le fonctionnement de la politique, et la situation en ce qui concerne l'exemption du paiement pour les enfants de moins de cinq ans au Ghana est exactement la même (Agyepong et Nagai, 2011). Le manuel d'exécution de la politique (Ministère de la Santé, 2006a) est resté dans les mains des seuls responsables de district après qu'ils l'eurent reçu en 2007 au démarrage. Ce manuel n'a pas été corrigé et diffusé de nouveau depuis. Pourtant, la définition des règles est primordiale dans le financement de la santé (Mathauer et Carrin, 2011). C'est ainsi que quelques agents peu scrupuleux, dans un système où la gouvernance est relativement faible, profitent du manque d'information ou de précision pour tirer des avantages du système. La situation est la même au Sénégal, au Ghana et au Niger (Agyepong et Nagai, 2011 ; Ridde et Diarra, 2009 ; Witter *et al.*, 2009 ; Witter *et al.*, 2010). Au Burkina Faso, si les agents avaient eu entre les mains le manuel (« *protocoles cliniques recommandés* », p. 32), ils auraient, par exemple, constaté qu'un accouchement peut être qualifié d'eutocique même si on réalise une épisiotomie et ils auraient donc demandé non pas 3 600 F CFA aux femmes mais plutôt 900 F CFA. L'affichage dans les maternités, deux ans après le début de la politique, des

tarifs du forfait n'ont pas suffi pour faire respecter ce dernier. En dehors des problèmes inhérents au contenu des kits, il faut signaler que les affiches sont en langue française, langue rarement comprise par les populations, et qu'aucun contrôle des dépenses réelles des femmes n'est effectué. Le Ministère a réalisé en 2009 la seule et unique enquête sur les dépenses déclarées par les agents de santé dans les registres, et les résultats de celle-ci n'avaient pas encore été rendus publics en 2011. Mais aucune enquête populationnelle n'a été réalisée. Comme ailleurs pour ce type de politique (Meessen *et al.*, 2009 ; Witter *et al.*, 2009), la fonction évaluative reste encore trop limitée.

Leçons apprises

Par-delà les leçons propres au contexte national de cette politique, la présente étude permet de dresser une liste d'un certain nombre de leçons apprises qui pourraient limiter l'écart de mise en œuvre de telles politiques *a priori* favorables à l'équité (encadré 1).

ENCADRÉ 1

Leçons apprises pour améliorer la mise en œuvre des politiques de subvention du paiement des soins

- Faire en sorte que les intervenants prennent part à la définition de la politique.
- Évaluer la mise en œuvre de la politique et ses outils avant de l'appliquer à l'échelle nationale.
- Mettre à jour régulièrement le manuel d'exécution de la politique en corrigeant les éventuelles erreurs et ambiguïtés.
- Rendre le contenu du manuel accessible et adapté à tous les niveaux du système de santé et à tous les agents de santé.
- Informer de manière continue la population sur le contenu et le fonctionnement de la politique.
- Effectuer régulièrement des sondages pour déterminer quels bénéfices le public cible tire de l'application de la politique.

Conclusion

Ce chapitre tâche de répondre à l'appel de Gilson et Raphaely (2008) pour « *more work on implementation, and specifically, the challenges of implementing equity-oriented policies* ». Nous avons donc usé d'une approche mixte de recherche qui nous a permis de tirer profit de la double force de l'approche qualitative et quantitative et ainsi de mieux mesurer et comprendre l'écart entre ce que les femmes déclarent payer pour un

accouchement et ce qu'elles devraient payer officiellement. Au cœur de l'explication se trouvent certainement le contenu institutionnel et les pratiques organisationnelles, car l'absence ou l'imprécision de certaines règles, des règles vagues de contrôle et la capacité organisationnelle sont des facteurs qu'il importe de prendre en compte pour arriver à une couverture universelle (Mathauer et Carrin, 2011).

Références

- AFRICAN UNION (2010), Assembly of the African Union. Fifteenth Ordinary Session, 25-27 juillet, Kampala, Ouganda.
- AGYEPONG, I. et R. A. NAGAI (2011), « "We charge them; otherwise we cannot run the hospital", front line workers, clients and health financing policy implementation gaps in Ghana », *Health Policy*, 99, 226-233.
- AMNESTY INTERNATIONAL (2009), *Giving Life, Risking Death. Maternal Mortality in Burkina Faso* (p. 112), Londres, Amnesty International Publications.
- BEN AMEUR, A. et al. (2010), *Perception de la qualité des soins et dépenses liées à l'accouchement par les parturientes à la suite de l'abolition du paiement des soins dans les districts sanitaires de Dori, Sebba et Djibo – Burkina Faso*. Ouagadougou, CRCHUM/CRDI/HELP/IRSS.
- BOOTH, D. (2011), « Towards a theory of local governance and public goods provision », *IDS Bulletin*, 42(2), 11-21.
- BRETON, E. et E. DE LEEUW (2011), « Theories of the policy process in health promotion research: a review », *Health Promotion International*, 26(1), 82-90.
- BRINKERHOFF, D. et B. Crosby (2002), *Managing Policy Reform: Concepts and Tools for Decision-Makers in Developing and Transitioning Countries*, West Hartford, Kumarian Press.
- CARROLL, C. et al. (2007), « A conceptual framework for implementation fidelity », *Implement Science*, 2, 40.doi:10.1186/1748-5908-2-40.
- CHEN, H. T. (2004), « Theory-driven evaluation », dans S. Mathison (dir.), *Encyclopedia of Evaluation* (p. 415-419), Thousand Oaks, Sage Publications.
- CRESWELL, J. W. et V. L. PLANO CLARK (2007), *Designing and Conducting Mixed Methods Research*, Thousand Oaks, Sage Publications.
- DE ALLEGRI, M. et al. (2011), « The impact of targeted subsidies for facility-based delivery on access to care and equity - Evidence from a population-based study in rural Burkina Faso », Communication présentée à la 3rd International Conference on Health Financing in Developing and Emerging Countries, 2011, 11-13 mai, Clermont-Ferrand, France, CERDI.
- DE BROUWERE, V. et W. VAN LERBERGHE (dir.) (2001), *Safe Motherhood Strategies: A Review of the Evidence*, Anvers, ITGPress.

- DIRECTION DES ÉTUDES ET DE LA PLANIFICATION (2007), *Annuaire statistique 2003 à 2007* (p. 165), Ouagadougou, Ministère de la Santé, Direction des études et de la planification.
- DURLAK, J. A. et E. P. DUPRE (2008), « Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation », *American Journal of Community Psychology*, 41(3-4), 327-350.
- GILSON, L. et N. RAPHAELY (2008), The terrain of health policy analysis in low and middle income countries: A review of published literature 1994-2007, *Health Policy and Planning*, 23, 294-307.
- GRAHAM, H., J. BELL et C. BULLOUGH (2001), « Can skilled attendance at delivery reduce maternal mortality in developing countries? », dans V. DE BROUWERE et W. VAN LERBERGHE (dir.), *Safe Motherhood Strategies: A Review of the Evidence* (vol. *Studies in Health Services Organisation and Policy*, p. 97-130), Anvers, ITG Press.
- MADON, T., K. J. HOFMAN, L. KUPFER et R. I. GLASS (2007), « Public health. Implementation science », *Science*, 318(5857), 1728-1729. doi:10.1126/science.1150009.
- MATHAUER, I. et G. CARRIN (2011), « The role of institutional design and organizational practice for health financing performance and universal coverage », *Health Policy*, 99, 183-192.
- MEESSEN, B. et al. (2009), *Removing User Fees in the Health Sector in Low-Income Countries: A Multi-country Review* (p. 37), New York, United Nations Children's Fund (UNICEF).
- MILES, M. B. et M. HUBERMAN (1994), *Qualitative Data Analysis: An Expanded Sourcebook* (2^e éd.), Newbury Park, CA, Sage Publications.
- MINISTÈRE DE LA SANTÉ (2006a), *Manuel d'exécution de la stratégie nationale de subvention des accouchements et des soins obstétricaux et néonataux d'urgence au Burkina Faso* (p. 45), Ouagadougou, Ministère de la Santé.
- MINISTÈRE DE LA SANTÉ (2006b), *Stratégie nationale de subvention des accouchements et des soins obstétricaux et néonataux d'urgence au Burkina Faso* (p. 65), Ouagadougou, Ministère de la Santé.
- NANDA, P. (2002), « Gender dimensions of user fees: Implications for women's utilization of health care », *Reproductive Health Matters*, 10(20), 127-134.
- OLIVIER DE SARDAN, J.-P. (2008), *La rigueur du qualitatif. Les contraintes empiriques de l'interprétation socio-anthropologique*, Louvain-la-Neuve, Academia Bruylant.
- OLIVIER DE SARDAN, J.-P. (2009), « State bureaucracy and governance in Francophone West Africa: An empirical diagnosis and historical perspective », dans G. BLUNDO et P.-Y. LE MEUR (dir.), *The Governance of Daily Life in Africa* (p. 39-72), Leyde, Brill.

- OLIVIER DE SARDAN, J.-P. (2010), « Le culturalisme traditionaliste africaniste. Analyse d'une idéologie scientifique », *Cahiers d'Études africaines*, L (2-3-4) (198-199-200), 419-453.
- OUATTARA, F., M.-E. GRUÉNAIS, S. ZONGO et R. OUEDRAOGO (2011), « Organiser l'activité de prévention de la transmission du VIH de la mère à l'enfant dans les structures périphériques : l'exemple du Burkina Faso », dans A. DESCLAUX, P. MSELLATI et K. SOW (dir.), *Les femmes à l'épreuve du VIH dans les pays du Sud. Genre et accès universel à la prise en charge* (p. 137-148), Paris, ANRS, coll. « Sciences sociales et SIDA ».
- PATTON, M. Q. (2008), *Utilization-Focused Evaluation* (4^e éd.), Thousand Oaks-Londres-New Delhi, Sage Publications.
- PLUYE, P. *et al.* (2011), *Proposal: A Mixed Methods Appraisal Tool for Systematic Mixed Studies Reviews*. Disponible sur : http://mixedmethodsappraisaltool-public.pbworks.com/f/MMAT_2011_criteria_and_tutorial_2011-06-29.pdf. Repéré [cité 22 août 2011].
- PRESSMAN, J. L. et A. WILDAVSKY (1984), *Implementation. How Great Expectations in Washington Are Dashed in Oakland* (3^e éd.), Berkeley, Los Angeles, Londres, University of California Press.
- RICHARD, F. *et al.* (2009), « The difficulty of questioning clinical practice: Experience of facility-based case reviews in Ouagadougou, Burkina Faso », *BJOG*, 116(1), 38-44.
- RICHARD, F., S. WITTER et V. DE BROUWERE (2010), « Innovative approaches to reducing financial barriers to obstetric care in low-income countries », *American Journal of Public Health*, octobre, 100, 1845-1852.
- RIDDE, V. (2008), « "The problem of the worst-off is dealt with after all other issues": The equity and health policy implementation gap in Burkina Faso », *Social Science and Medicine*, 66, 1368-1378.
- RIDDE, V. (2011), « Politiques publiques de santé, logiques d'acteurs et ordre négocié au Burkina Faso », *Cahiers d'Études africaines*, 1(201), 115-143.
- RIDDE, V. et A. DIARRA (2009), « A process evaluation of user fees abolition for pregnant women and children under five years in two districts in Niger (West Africa) », *BMC Health Services Research*, 9(89).
- RIDDE, V. et F. MORESTIN (2011), « A scoping review of the literature on the abolition of user fees in healthcare services in Africa », *Health Policy and Planning*, 26(1), 1-11.
- RIDDE, V. *et al.* (2011), « The national subsidy for deliveries and emergency obstetric care in Burkina Faso », *Health Policy and Planning*, Supplement 2(26), ii30-ii40.
- SAETREN, H. (2005), « Facts and myths about research on public policy implementation: Out-of-fashion, allegedly dead, but still very much alive and relevant », *Policy Studies Journal*, 33(4), 559-582.

- SOMBIÉ, I. *et al.* (2007), *Évaluation de la mise en place de la subvention des accouchements et des soins obstétricaux et néonataux d'urgence au Burkina. Rapport préliminaire* (p. 60), Ouagadougou, Ministère de la Santé, GREFsAD.
- UNICEF (2009), *The State of the World's Children 2009. Maternal and Newborn Health*. New York, United Nations Children's Fund (p. 158), New York, UNICEF.
- UNITED NATIONS SECRETARY-GENERAL (2010), *Global Strategy for Women's and Children's Health* (p. 20), New York, United Nations.
- WALKER, L. *et* L. GILSON (2004), « "We are bitter but we are satisfied": Nurses as street-level bureaucrats in South Africa », *Social Science and Medicine*, 59(6), 1251-1261.
- WALT, G. *et al.* (2008), « "Doing" health policy analysis: Methodological and conceptual reflections and challenges », *Health Policy Plan*, 23(5), 308-317.
- WITTER, S. *et* S. ADJEI (2007), « Start-stop funding, its causes and consequences: A case study of the delivery exemptions policy in Ghana », *Int J Health Plann Manage*, 22(2), 133-143.
- WITTER, S., S. ADJEI, M. ARMAR-KLEMESU *et* W. GRAHAM (2009), « Providing free maternal health care: Ten lessons from an evaluation of the national delivery exemption policy in Ghana », *Global Health Action*. Publié en ligne le 13 janvier 2009.
- WITTER, S. *et al.* (2010), « The national free delivery and caesarean policy in Senegal: Evaluating process and outcomes », *Health Policy and Planning*, septembre, 25(5), 384-92.
- WORLD HEALTH ORGANIZATION (2010a), *Countdown to 2015 Decade Report (2000-2010): Taking Stock of Maternal, Newborn and Child Survival* (p. 56), Genève, World Health Organization *et* United Nations Children's Fund.
- WORLD HEALTH ORGANIZATION (2010b), *The World Health Report-Health Systems Financing: The Path to Universal Coverage* (p. 106), Genève, World Health Organization.
- WORLD HEALTH ORGANIZATION (WHO), United Nations Children's Fund (UNICEF), United Nations Population Fund (UNFPA) *et* World Bank (2010), *Trends in Maternal Mortality: 1990 to 2008* (p. 55), Genève.
- YIN, R. K. (2009), *Case Study Research: Design and Methods* (4^e éd.), Los Angeles, Sage Publications.

Une évaluation d'implantation et d'efficacité d'un programme destiné aux enfants et à leur famille

*Christian Dagenais, Frédéric Nault-Brière,
Didier Dupont et Julie Dutil*

Le réseau des services mis en place à l'intention des jeunes à travers le monde fait face à des demandes de plus en plus complexes et dispose de ressources relativement limitées (Reitan, 1998; Scott, 2005). La nécessité d'une coordination efficace entre les organismes de services n'a jamais paru aussi pressante, mais les forces d'inertie pèsent sur le travail en collaboration (D'amour *et al.*, 2004; Mattesich, Murray-Close et Monsey, 2001; Wildridge, Childs, Cawthra et Madge, 2004). Des obstacles se dressent à plusieurs niveaux, tels le manque de clarté dans la définition des rôles et des responsabilités (Wildridge *et al.*, 2004), l'insuffisance des ressources et l'absence de budgets communs (Sloper, 2004), le peu de temps consacré à la collaboration et le nombre excessif de cas (Harker, Dobel-Ober, Berridge et Sinclair, 2004; Sloper, 2004). De plus, la prédominance d'un modèle professionnel dans lequel les travailleurs cherchent à obtenir l'autonomie afin de préserver leurs attributions (Abbott, 1988; San Martin-Rodriguez, Beaulieu, D'Amour et Ferrada-Videla, 2005)

empêche l'établissement de relations interorganisationnelles fluides. Dans ce contexte, l'identification de modèles efficaces de coordination présentant un bon rapport coût-efficacité représente, à la fois, un objectif central et un enjeu vital des réseaux publics.

Contexte et description du programme

Afin d'améliorer la continuité des interventions, le ministère de la Santé et des Services sociaux (MSSS) du Québec a créé en 2002 une stratégie d'action à l'intention des jeunes en difficulté et de leur famille aux prises avec des problèmes complexes. Un des éléments centraux de cette stratégie était la mise sur pied d'équipes de coordination des services, appelées Équipes d'intervention jeunesse (ÉIJ), dans chacune des 17 régions administratives de la province.

Le projet pilote consistait à former au moins une ÉIJ dans chacune des régions. Ces équipes interdisciplinaires ont été mandatées pour faciliter l'accessibilité et la coordination des services dans des situations particulièrement complexes impliquant des jeunes (0-18 ans). Les ÉIJ sont composées d'un coordonnateur responsable du fonctionnement de l'équipe et d'agents de liaison qui représentent chacun un partenaire du réseau jeunesse. Les coordonnateurs sont chargés d'assurer la liaison entre l'équipe et les ressources disponibles dans leurs établissements respectifs. Les partenaires visés sont les Centres de santé et de services sociaux (CSSS)¹, les centres jeunesse², les établissements en déficience physique et intellectuelle, les ressources en santé mentale et en toxicomanie ainsi que les partenaires jugés pertinents, tel le milieu scolaire. La plupart des établissements partenaires ont désigné un cadre intermédiaire pour les représenter. Les ÉIJ devaient prendre en charge des situations qui concernent plusieurs établissements et devaient prioriser (a) les situations considérées comme étant dans une impasse clinique ou administrative et (b) les situations qui ne font pas clairement partie du mandat des établissements.

Les interventions des ÉIJ consistaient en une action structurée qui pouvait s'échelonner sur quelques jours ou quelques mois. Les cas référés

1. Les CSSS regroupent les établissements qui offrent les soins de santé et les services sociaux à la population.

2. Les centres jeunesse regroupent les établissements qui fournissent des services aux jeunes en besoin de protection et sous le coup de la *Loi sur le système de justice pénale pour adolescents*.

par les agents de liaison ont été évalués conjointement par tous les membres de l'équipe, puis sélectionnés selon les critères définis dans les lignes directrices du programme. Lorsqu'une situation était acceptée, les ÉIJ déterminaient les services dont les jeunes avaient besoin et l'agent de liaison désignait les intervenants des établissements qui participeraient au processus clinique. Chaque cas nécessitait le concours de plusieurs intervenants, et certains prenaient part aussi à plusieurs interventions de l'ÉIJ. Le processus d'intervention s'amorçait par une rencontre entre la famille de l'enfant, les intervenants des établissements désignés et le coordonnateur de l'ÉIJ pour préparer un plan de services individualisés (PSI). Le PSI est un outil qui vise à assurer la coordination et la continuité des services dispensés aux bénéficiaires ainsi que leur participation active à la définition de leurs besoins et des interventions à réaliser. Un des intervenants était ensuite désigné comme intervenant principal et mandaté pour assurer une liaison constante entre les familles et les intervenants entre les réunions de préparation des PSI et après le retrait de l'ÉIJ.

Plan d'évaluation

L'évaluation visait trois objectifs : 1) déterminer le degré d'implantation des ÉIJ en fonction de ce qui était prévu au départ ; 2) documenter les effets des ÉIJ sur les enfants et leur famille ainsi que sur les organisations participantes ; et 3) évaluer les processus de coordination mis en place dans les équipes. Ce dernier objectif consistait à cerner et à analyser les facteurs qui influencent l'implantation des équipes et l'efficacité des services de coordination offerts afin d'expliquer les résultats.

L'approche retenue pour atteindre ces objectifs repose sur une méthode mixte alliant des techniques de groupes et individuelles ainsi que sur des analyses quantitatives et qualitatives. La mise en commun des différentes perspectives, obtenues par la triangulation de diverses méthodes, permet de déterminer les convergences entre les résultats et d'augmenter leur crédibilité. Cette stratégie fait également en sorte que les biais inhérents à une méthode particulière peuvent être compensés par la mise à contribution d'une autre (Campbell et Fiske, 1959 ; Cook, 1985). Ainsi, trois sources de données complémentaires ont été utilisées : une banque de données quantitatives (en ligne), des entretiens en profondeur réalisés

dans quatre régions et des études de cas des familles. Une synthèse de la méthodologie est présentée au tableau 22.1 et le processus d'évaluation est décrit dans la section qui suit.

TABLEAU 22.1
Synthèse de la méthodologie mixte

Méthode de collecte de données	Échantillon	Instruments	Stratégie d'analyse	Objectifs d'évaluation
Banque de données quantitatives	441 situations provenant des 17 équipes	Formulaire en ligne rempli par les coordonnateurs des ÉIJ	Statistiques descriptives	Implantation (1)
Études de cas	29 cas provenant de 13 équipes (114 entrevues)	Entrevues semi-structurées	Analyse thématique	Impacts (2) Processus de coordination (3)
Analyse approfondie de quatre régions	56 répondants en individuel et 4 <i>focus groups</i> (n = 37) lors de réunions d'équipe	Entrevues semi-structurées et <i>focus groups</i>	Analyse thématique	Impacts (2) Processus de coordination (3)

Note: Certains participants ont été interviewés à la fois pour les études de cas et l'analyse des quatre équipes.

Description du processus d'évaluation

Constitution de l'échantillon

La banque de données en ligne comprenait des données quantitatives détaillées sur chaque situation prise en charge par les ÉIJ. Les données ont été fournies par les coordonnateurs des ÉIJ. Ces équipes ont pris en charge 441 situations qui leur ont été adressées.

Pour les études de cas, une stratégie en deux étapes a été conçue pour mettre sur pied un échantillon de 30 situations représentatives et théoriquement significatives. Toutes les situations ont d'abord été classées dans une table de sélection aléatoire stratifiée selon quatre critères : la date de la référence à l'équipe, le groupe d'âge du jeune (10 situations par groupe d'âge), le référent (10 situations pour chacun des trois référents principaux : CLSC, milieu scolaire et centre jeunesse) et la présence d'un PSI (50 % des

situations avec PSI et 50 % sans PSI). Ensuite, les situations ont été choisies au hasard dans chaque catégorie de la table. Pour chaque cas sélectionné, des entrevues individuelles ont été menées avec (a) le coordonnateur des ÉIJ, (b) l'agent de liaison qui a adressé le cas, (c) l'intervenant principal et (d) la famille ayant reçu les services. En tout, 13 des 17 ÉIJ ont été représentées dans les études de cas, et chacune avait entre une et six situations.

Quatre équipes ont été sélectionnées en concertation avec le comité de liaison du projet, et des analyses plus approfondies ont été réalisées dans les villes suivantes : Rivière-du-Loup, Sept-Îles et Gatineau, ainsi que l'arrondissement Montréal-Nord. Sur une période de deux à quatre jours, des membres de l'équipe d'évaluation ont rencontré les coordonnateurs, les supérieurs immédiats, les agents de liaison et les répondants jeunesse de chacune de ces équipes. À des fins de triangulation, chacune des quatre équipes a participé à un groupe de discussion pour compléter l'analyse approfondie. Le groupe comprenait le coordonnateur, les agents de liaison et, dans un cas, les répondants jeunesse. Au total, 174 entrevues ont été réalisées individuellement (170) et en groupe (4 regroupant un total de 37 intervenants).

Instruments

Les coordonnateurs ont rempli un formulaire en ligne pour chacune des situations traitées. Les données portent sur l'âge du jeune, les problématiques identifiées, les organismes concernés, les raisons pour lesquelles le cas a été soumis et la présence ou non d'un PSI. Le contenu des fiches et des statistiques fournies par les coordonnateurs a été systématiquement vérifié et validé auprès d'eux afin d'éliminer les données contradictoires, incomplètes ou erronées (une erreur de date, par exemple). Cet instrument a été mis au point par l'équipe d'évaluation en étroite collaboration avec quatre coordonnateurs d'ÉIJ.

Des entrevues semi-structurées ont été utilisées pour les études de cas et l'analyse approfondie des équipes de quatre régions. Elles ont été conçues par l'équipe d'évaluation et pré-testées par les coordonnateurs et les agents de liaison. Les entrevues des études de cas duraient entre 30 et 45 minutes et portaient sur : 1) la description de la situation et son évolution ; 2) les critères qui ont mené le cas à être soumis ; 3) la nature des services fournis ; 4) les obstacles rencontrés ; 5) l'impact perçu de

l'intervention; et 6) les sources de satisfaction et d'insatisfaction des participants. Le niveau de satisfaction des familles a été déterminé de façon systématique.

Les entrevues individuelles menées en vue de l'analyse approfondie des quatre régions ont duré entre 30 et 90 minutes et touchaient des sujets tels que les rôles joués par les différents partenaires, les défis perçus du travail en collaboration, les critères utilisés pour définir une situation des ÉIJ, les facteurs qui ont facilité ou empêché l'élaboration des PSI et la manière dont les ressources financières ont été utilisées. Les groupes de discussion ont duré entre 60 et 90 minutes. Trois thèmes généraux ont été abordés: 1) l'implication des agents de liaison dans les ÉIJ; 2) les avantages de la coordination; et 3) les défis liés à la réalisation d'un PSI.

Collecte des données

Les coordonnateurs ont rempli les formulaires de la banque de données fournis sur un site internet accessible seulement avec un code d'identification personnalisé. Ils avaient au préalable demandé l'autorisation des parents. Les rencontres avec ces derniers avaient été planifiées avec l'aide des coordonnateurs, qui étaient bien placés pour obtenir ce consentement parental. En ce qui concerne les études de cas cliniques, des cas choisis aléatoirement ont été remis aux coordonnateurs seulement deux semaines avant les entrevues pour éviter les biais (par exemple, choisir les situations dans lesquelles les équipes se sont beaucoup investies). Les entrevues ont d'abord été enregistrées sur bande magnétique, puis transcrites. L'équipe d'évaluation s'est également chargée de faire respecter une politique stricte de confidentialité et a obtenu l'approbation éthique de l'Université de Montréal.

Analyses des données

Les données quantitatives de la banque de données en ligne ont été analysées à l'aide de statistiques descriptives telles que les moyennes et les fréquences. L'organisation, la codification et l'analyse de l'ensemble du matériel qualitatif ont été réalisées à l'aide du logiciel de traitement de données qualitatives Nvivo 2 (QSR). Le contenu des entrevues a d'abord été structuré en fonction des objectifs deux et trois du plan d'évaluation (effets et processus de coordination). Le matériel a ainsi été structuré autour

de thèmes que l'équipe d'évaluation trouvait importants d'analyser, mais aussi autour de ceux qui sont apparus au fil de l'analyse des discours recueillis. Cette technique est inspirée de la *grounded theory*, développée par Glaser et Strauss (1967) et Strauss et Corbin (1990). Dans un premier temps, on s'est attaché à déceler les concepts sous-jacents ou les thèmes qui caractérisent les meilleurs extraits du matériel discursif recueilli. Chaque fois qu'un thème nouveau et pertinent émergeait, un code était créé dans le logiciel. Les extraits liés aux concepts ou thèmes déjà répertoriés ont été stockés sous les codes existants, ce qui permettait d'ajouter de la profondeur à ces thématiques. Les codes ont ensuite été liés schématiquement et organisés selon des catégories plus larges. Par exemple, tous les codes se rapportant aux relations entre les membres de l'ÉIJ ont été groupés sous le thème de la « collaboration interprofessionnelle ». À l'étape de l'écriture, des confrontations constantes entre les données et les conclusions ont permis de s'assurer que les conclusions étaient conformes au matériel qualitatif. Lorsque cela était possible, les mots des participants ont été utilisés pour soutenir les thématiques qui émergeaient ainsi que les conclusions. Pour assurer la triangulation, les résultats ont été validés systématiquement par plusieurs membres de l'équipe (Patton, 1990).

L'évaluation de l'implantation des ÉIJ (objectif 1) reposait sur une comparaison des données quantitatives obtenues à partir de la banque de données et tenait compte principalement de deux dimensions. Le MSSS avait déterminé les cibles à atteindre avant l'implantation : les équipes devaient prendre en charge entre 30 et 40 situations par année et préparer un PSI pour 75 % d'entre elles.

L'évaluation des effets des ÉIJ (objectif 2) était basée sur des données qualitatives provenant des études de cas et de l'analyse approfondie des équipes de quatre régions. Les perceptions des participants et spécialement des membres de la famille à l'égard de l'efficacité des interventions des ÉIJ ont été analysées. Les changements de perception quant à l'efficacité et à la satisfaction avant et après les interventions ont été considérés comme particulièrement instructifs. Comme nous l'expliquerons plus loin, le choix de la méthode d'évaluation qualitative pour analyser les impacts du programme a été largement dicté par des facteurs contextuels. L'évaluation des processus de coordination (objectif 3) reposait essentiellement sur des données qualitatives, mais des données quantitatives ont aussi été utilisées quelques fois pour appuyer les conclusions.

Résultats

Les résultats sont présentés séparément pour chacun des objectifs de l'évaluation.

Objectif 1: implantation des ÉIJ

Clientèle concernée

Les 17 équipes d'intervention jeunesse ont traité en moyenne 25,9 situations ($N = 441$; $SD = 14,1$) au cours de la seconde année du projet. Ces résultats montrent une nette amélioration par rapport à la moyenne de la première année ($M = 19,6$; $N = 333$; $SD = 15,7$) et s'approchent de la cible de 30 fixée par le MSSS. Cependant, le nombre de situations varie considérablement d'une équipe à l'autre; certaines d'entre elles ont traité 5 situations et d'autres jusqu'à 53.

Les résultats montrent qu'aucun groupe d'âge n'a été surreprésenté ou sous-représenté parmi les situations traitées par les ÉIJ. Ils montrent également que, comme prévu, les CLSC, les centres jeunesse et les écoles ont été de loin les établissements qui ont adressé le plus de cas. Ensemble, ils représentent près de 75 % du nombre total des situations adressées. Aucun des autres organismes (centres de réadaptation en déficience intellectuelle et physique, garderies, organismes communautaires, etc.) n'a adressé plus de 7 % des situations. Les cas des ÉIJ comportaient pour la plupart de multiples problématiques avec une moyenne de 3,1 ($N = 441$; $SD = 1,5$) problématiques par situation (troubles du comportement, abus physique, violence psychologique, négligence, abus sexuels, etc.). La plupart du temps, les situations impliquent deux (25,2 %) ou trois (29,5 %) problématiques. Les problèmes parentaux (50,8 %) et les problèmes de conduite de l'enfant (49,4 %) ont été les plus fréquemment observés. Se fondant sur ces données, et même si une cible n'avait pas été préalablement fixée, le MSSS a jugé que les ÉIJ avaient rejoint la clientèle visée.

Participation des partenaires

Le nombre moyen d'établissements impliqués dans les situations était de 3,6 ($N = 441$; $SD = 1,1$). Plus précisément, 0,9 % des situations impliquaient une seule agence, 17,0 % en mobilisaient deux, 65,3 % trois ou quatre, et

16,8 % étaient confiées à cinq agences ou plus. Selon le MSSS, les résultats obtenus autorisaient à conclure que les ÉIJ avaient produit les interventions interorganisationnelles attendues.

Plans de services individualisés (PSI)

La proportion globale de PSI durant la deuxième année, y compris ceux qui étaient en préparation, a été de 78,66 % (N = 417), ce qui correspondait au critère fixé par le MSSS (75 %). Vingt-quatre situations ont été exclues des analyses parce que le PSI n'a pas été jugé nécessaire pour différentes raisons par l'équipe d'évaluation (par exemple, une seule agence était impliquée avec la famille). Ce résultat représente une amélioration considérable par rapport à la première année, car, durant celle-ci, les PSI avaient été complétés dans 56,8 % des cas.

Objectif 2 : les effets des ÉIJ

Les résultats des équipes ont été évalués en se basant exclusivement sur l'analyse du contenu qualitatif provenant des entrevues menées avec les professionnels des ÉIJ (études de cas et analyse approfondie de quatre équipes) et des familles (études de cas). Les sections suivantes présentent les résultats de ces analyses.

Effets des services de coordination sur les enfants et leur famille

Les membres des familles impliqués dans les situations des ÉIJ étaient pour la plupart satisfaits du travail accompli par les équipes. Les rares insatisfactions exprimées portaient sur le réseau de services et étaient relatives aux listes d'attente et aux retards organisationnels plutôt qu'au fonctionnement de l'équipe. Bien que les familles n'aient pas toujours été entièrement satisfaites des résultats, des procédures et des interventions, toutes ont apprécié les efforts déployés par les coordonnateurs. Dans l'ensemble, la satisfaction a été plus souvent liée au processus (relations avec les professionnels, organisation des rencontres, etc.) qu'aux résultats (qualité et accessibilité des services, amélioration clinique des enfants).

Globalement, les familles et les intervenants ont noté que la coordination des ÉIJ a eu des effets concrets sur les services offerts. L'amélioration

de l'accès et la qualité des services ont été les facteurs les plus souvent nommés. En ce qui concerne la qualité, les familles ont souvent remarqué que la coordination amène une meilleure organisation des services et des interventions plus adaptées aux besoins et aux capacités des familles. Cela démontre que les équipes ont fourni des services efficaces et centrés sur le client.

Bien que l'impact sur les enfants n'ait pas été directement évalué à l'aide d'outils standardisés, l'analyse des observations des membres des familles et les perceptions au sujet du bien-être de leur enfant suggèrent que les changements survenus dans les services après la coordination des ÉIJ ont eu des effets positifs. Ces changements ont été également bénéfiques pour les autres membres de la famille. Les parents épuisés, pessimistes et désillusionnés par les expériences passées se sont souvent sentis soulagés à la suite des interventions des ÉIJ. Ces interventions ont contribué à restaurer la confiance des parents envers le réseau et à renforcer les liens entre les familles et les professionnels. Le pouvoir d'agir (*empowerment*) des parents a également augmenté puisque les membres de la famille ont pu jouer un rôle central dans le processus de collaboration. Les parents ont apprécié le fait qu'ils étaient respectés et considérés comme des membres de l'équipe. Voici quelques exemples des commentaires provenant des familles :

Oui. Je trouve que ça m'en enlève beaucoup, parce qu'il y a beaucoup de gens qui vont prendre des choses et ça se fait une fois et tout le monde a compris la même chose. Donc, je n'ai pas à répéter. Souvent, avant, ce qui arrivait, c'est que je devais répéter mon histoire à l'un et à l'autre et souvent une personne pouvait avoir mal compris, et c'est là que ça tournait en rond. Tandis que là, tout le monde part avec l'idée et chacun s'enlève sur quelque chose. Et [le coordonnateur] va rappeler pour voir si tout est correct et où en est rendue telle demande et tout ça. Fait que ça aussi, moi, ça m'en enlève beaucoup. – Une famille

Parce que moi, dans le fond, avant, il y avait comme un manque de crédibilité. Il y en a qui veulent abuser du système, ça fait que là, ils avaient toujours cette crainte-là. Ça fait que moi, je sentais ça beaucoup avant. Et là, le fait que j'ai du monde qui vient dire et que ce que je dis est proche de la réalité, à ce moment-là, on devient plus crédible. Et ça enlève toute une attitude au niveau de la table, où les gens essaient plus d'aider et de trouver une solution plutôt que de laisser ça de même. – Une famille

Effets sur les pratiques et les organisations

Les intervenants, tout comme les membres de la famille, ont été influencés de manière positive par la coordination des ÉIJ. Plusieurs répondants mentionnent les avantages de la responsabilité partagée et se rendent compte qu'ils ne sont plus les seuls responsables du dossier. Chacun assume les responsabilités qu'il a acceptées au départ. La coordination a aussi renforcé les liens entre les intervenants et elle a même souvent permis d'apaiser certaines tensions. Ils ont ainsi acquis le sentiment de former une équipe et ils se sentent moins seuls pour gérer des situations complexes.

Je pense que la part de l'ÉIJ, dans ce contexte-là, c'est d'être capable de raviver la flamme et de donner un sens commun, donc de permettre aux gens d'adhérer à un groupe aussi, de briser l'isolement. Ça fait que ça permet à l'engagement de revenir s'il n'était plus là ou, au moins, de s'actualiser dans quelque chose de très concret. Ça fait que je pense que, dans ce sens-là, l'ÉIJ a un apport extraordinaire. – Un agent de liaison

Et la mise en commun de nos expertises, de nos services, tout ça, concentré avec le parent, fait en sorte que c'est facilitant. C'est pour ça qu'un dossier complexe comme ça n'aurait pas pu se faire avec la méthode traditionnelle d'entente mutuelle. C'est très clair pour moi. Parce qu'il y a eu des moments de frustration tels qu'il fallait au moins sentir qu'on n'était pas tout seuls là-dedans. Ça faisait baisser les tensions générées par les frustrations. [...]. Mais moi, dans l'ensemble, j'ai toujours eu l'impression qu'on se respectait et qu'on s'engageait du mieux qu'on pouvait dans les dossiers. Ça fait que ça, ça a renforcé l'adhésion que, moi, j'avais à l'ÉIJ. – Un agent de liaison

Les résultats des analyses montrent qu'en plus de l'effet immédiat que les ÉIJ ont sur les situations, les équipes produisent un effet indirect à long terme sur les pratiques et les organisations. Cet effet est particulièrement évident avec l'utilisation des plans de services individualisés. La réalisation de PSI représente un problème important dans le réseau pour la simple raison qu'ils ne sont que très rarement utilisés par les intervenants même s'ils sont requis par la loi dans tous les cas interdisciplinaires (Brière *et al.*, 2007). Selon les répondants, les ÉIJ ont fait plus que simplement initier et compléter des PSI dans les situations complexes, elles ont également contribué à introduire une culture PSI. Les travaux des équipes ont démontré que les PSI étaient utiles et efficaces, et cela a incité les

professionnels à les utiliser davantage. Un autre impact indirect à long terme porte sur l'établissement de relations informelles entre les professionnels et sur les liens entre les outils d'intervention qui peuvent aussi être utilisés dans des situations autres que celles dont les ÉIJ ont à s'occuper.

Ce qui est intéressant aussi, vous savez que, depuis plusieurs années, on est beaucoup formés dans l'interdisciplinarité et que c'est dur à vivre au quotidien. Et je dois dire qu'ici [à l'ÉIJ], c'est vraiment une rampe de lancement pour apprendre à travailler dans l'interdisciplinarité. On a tous des fonctions et des professions différentes, on provient tous d'établissements différents, avec des objectifs différents. Pour le bien-être des enfants et de leur famille, on apprend à mettre en commun l'interdisciplinarité. Et c'est très rassurant. Je trouve que professionnellement, ça fait beaucoup de bien de venir ici et de passer par ici.

– Un agent de liaison

Objectif 3 : Processus de coordination

La très grande majorité des répondants ont jugé que la structure et les ressources allouées à l'ÉIJ étaient adéquates. Ce qui a été grandement apprécié des participants est le lieu officiel qui leur avait été attribué et le temps dont ils disposaient pour échanger et coordonner les services. Certains participants ont souligné que le contexte structuré et bien défini du programme favorisait davantage la collaboration que les lignes directrices informelles généralement utilisées dans le réseau. Les intervenants ont aussi fréquemment mentionné que le fait d'avoir des agents de liaison a rendu les ÉIJ plus efficaces que les autres modèles de coordination. Ces agents disposent du pouvoir de surmonter les obstacles organisationnels et ils peuvent effectuer les changements qui s'imposent pour que le client reçoive les services requis dans des situations complexes. Les professionnels doivent souvent respecter les limites rigides imposées par leur organisation.

Ensuite, les intervenants et les familles considèrent que l'introduction d'un coordonnateur indépendant, expert et légitime a été la principale innovation apportée par le programme. Les coordonnateurs ont joué un rôle central dans le fonctionnement administratif et clinique de leurs équipes, et leur présence a été particulièrement utile pour animer les processus durant les rencontres PSI.

J'ai vécu tellement d'expériences où ça n'a pas fonctionné justement parce qu'il n'y avait pas de coordination. Ce qui est bien, c'est que la coordination travaille pour les familles mais aussi pour nous autres. Nous rassembler... Elle s'occupe de l'organisation, facilite les choses. Et son but, c'est la famille, mais avec les particularités de chaque établissement. Au début, je me demandais pourquoi la coordonnatrice ÉIJ mettait tant l'accent sur ce qu'on allait faire ensemble et ce qui nous unissait... Et là, maintenant, je comprends. Parce que si on garde le focus juste sur notre rôle, on va revenir avec les mêmes problèmes qu'on avait avant, c'est-à-dire que chacun travaille dans sa petite case et qu'on n'a pas de chemin commun. Avec la coordination, notre chemin, c'est la famille, et on arrive à un but et à un bien-être, à sortir des problématiques. – Un agent de liaison

Enfin, la qualité de la collaboration entre les partenaires a été un déterminant central du fonctionnement de l'équipe. Quatre processus de collaboration importants ont caractérisé la dynamique interne des équipes les plus efficaces : 1) la formulation et la poursuite de buts communs ; 2) le maintien de bonnes relations interpersonnelles ; 3) le développement et la consolidation de l'adhésion des partenaires ; et 4) une formalisation appropriée des règles de fonctionnement de l'équipe. La collaboration au sein d'une équipe a été influencée par des facteurs qui variaient d'une région à l'autre : les relations entre les organismes avant l'implantation du programme, l'engagement de l'agence, la structure de l'équipe et les ressources dont elle disposait.

Discussion

Les résultats de l'évaluation montrent que les ÉIJ étaient globalement bien établies. Les attentes concernant la clientèle desservie et la nature des interventions prévues ont également été satisfaites. Les situations traitées par les ÉIJ comportaient plusieurs problématiques et étaient également réparties entre les groupes d'âge. De plus, les interventions impliquaient généralement plusieurs établissements, et un plan de services individualisés a été la plupart du temps élaboré. Cependant, le nombre moyen de cas pris en charge (25,9) était légèrement inférieur au nombre ciblé (30). Cette différence est sans doute attribuable au fait que certaines équipes existaient depuis moins d'un an et donc qu'elles n'utilisaient pas encore toutes leurs capacités. Le fait que la plupart des équipes établies durant la première année ont pris en charge plus de situations durant la deuxième année vient appuyer cette hypothèse.

L'évaluation a également montré que le programme produit les effets escomptés. Les familles et les professionnels ont été extrêmement satisfaits des interventions des ÉIJ. Les insatisfactions étaient rares, légères et n'étaient généralement pas liées au fonctionnement de l'équipe. De plus, les ÉIJ ont contribué à l'établissement d'une culture pro-PSI. L'évaluation des processus de coordination a permis de déterminer et aussi de comprendre les facteurs qui ont influé sur l'implantation du programme et contribué à le rendre efficace.

Considérations méthodologiques

Les évaluations des équipes d'intervention jeunesse comportent des défis méthodologiques complexes. Des choix ont dû être adaptés au contexte de l'évaluation. Les ÉIJ font partie du projet pilote étalé sur trois ans (2003-2006) et les décisions concernant la poursuite du programme ont été prises six mois avant la fin de la troisième année. Étant donné que l'implantation d'un programme prend du temps (Rossi, 1992a, 1992b), apporter des conclusions solides sur le programme et ses impacts dans un laps de temps si court a été un défi considérable pour l'équipe d'évaluation. Les méthodes quantitatives exhaustives et de longue haleine ont donc été délaissées au profit de techniques plus flexibles. Les essais contrôlés randomisés (ECR) qui, selon certains, permettent de déterminer de la façon la plus valide scientifiquement les impacts du programme n'ont pu être utilisés. L'absence d'une clientèle cible bien définie, essentielle pour l'établissement d'un groupe contrôle, est à l'origine de cette décision.

Étant donné ce contexte, une mesure qualitative des perceptions des différentes parties intéressées a été utilisée pour évaluer l'efficacité du programme. Bien que les méthodes basées sur les perceptions comportent des limites importantes (Worthington, 2005), la validité des résultats présentés dans ce chapitre est démontrée par plusieurs éléments. Les parents ont pu exprimer leurs opinions sur les services qu'ils ont reçus, de telle sorte que l'efficacité du programme a été évaluée en partie à partir de leur point de vue et en fonction de leur définition de l'efficacité (qui est au moins aussi importante que la définition du prestataire de services). En outre, les questions ont été posées aux parents de manière systématique, ce qui a permis de conclure que la perception qu'ont les parents de l'efficacité des services de coordination a été étroitement liée aux actions

des ÉIJ. Étant donné que les parents et les professionnels portaient un jugement presque unanimement positif sur le programme, il est raisonnable de conclure que l'évaluation fournit des preuves robustes de l'efficacité des ÉIJ.

Un certain nombre de techniques ont été utilisées pour assurer et améliorer la validité de l'évaluation. La triangulation (Patton, 1990) était une de ces techniques. Les perceptions de différents participants ont été obtenues pour chaque équipe et des études de cas ont été réalisées (triangulation des sources), trois méthodes de collecte de données ont été employées (triangulation des méthodes) et les analyses ont été réalisées et validées par différents chercheurs (triangulation des chercheurs). Cette stratégie a compensé les biais inhérents à toute méthode d'évaluation unique (voir à ce sujet le chapitre 9 de Bamberger et Rugh dans cet ouvrage; Cook et Campbell, 1979; Cook, 1985) et des données de plusieurs perspectives complémentaires ont permis une meilleure compréhension du problème de recherche (Chenitz et Swanson, 1986; Glaser, 1978; Posavac et Carey, 1992; Strauss, 1987). Enfin, les participants ont présenté et validé les résultats préliminaires pour améliorer la validité écologique³ de l'évaluation (Fawcett, 1990; Tandon, Azelton, Kelly et Strickland, 1998).

Conclusion

Ce chapitre démontre que les décisions méthodologiques doivent prendre en compte le contexte politique et pratique d'un programme. Les limites temporelles et les contextes d'évaluation mal définis imposent le choix de stratégies adaptables et souples qui peuvent être toujours valides écologiquement à condition d'inclure des techniques reconnues comme la triangulation (Patton, 1990). L'intégration des méthodes qualitatives et quantitatives dans un modèle d'évaluation basé sur les perceptions des participants et des intervenants peut offrir une stratégie intéressante pour obtenir des résultats solides et pertinents.

3. La validité écologique est le niveau selon lequel l'expérience vécue par le sujet d'une expérimentation correspond à celui que suppose le chercheur (Bronfenbrenner, 1977).

Références

- ABBOTT, A. (1988), *The System of Professions: An Essay on the Division of Expert Labour*, Chicago, University of Chicago.
- BRIÈRE, F. et al. (2007), « Relever le défi des plans de services individualisés (PSI) au Québec: leçons tirées de l'expérience des équipes d'intervention jeunesse (ÉIJ) », *Revue canadienne de santé mentale communautaire*, 26(1), 97-112.
- BRONFENBRENNER, U. (1977), « Toward an experimental ecology of human development », *American Psychologist*, 32, 513-531.
- CHENITZ, W. et J. SWANSON (1986), *From Practice to Grounded Theory: Qualitative Research in Nursing*, Menlo Park, Addison-Wesley.
- CHERTOK, F. et L. JACKSON (dir.) (1990), *Researching Community Psychology* (p. 64-75), Washington, D.C., American Psychological Association.
- COOK, T. D. (1985), « Post-positivist critical multiplism », dans L. SHOTLAND et M. MARK (dir.), *Social Science and Social Policy*, Beverly Hills, Sage.
- COOK, T. D. et D. T. CAMPBELL (1979), *Quasi-experimentation*, Chicago, Rand McNally.
- D'AMOUR, D., M. D. BEAULIEU, L. SAN MARTIN RODRIGUEZ et M. FERRADA-VIDELA (2004), « Key elements of collaborative practice and frameworks: Conceptual basis for interdisciplinary practice », dans *Interdisciplinary Education for Collaborative, Patient-Centred Practice: Research and Findings Report* (p. 63-89), Ottawa, Santé Canada.
- FAWCETT, S. (1990), « Some emerging standards for community research and actions: Aid from a behavioral perspective », dans P. TOLAN et F. C. KEYS (dir.).
- GLASER, B. G. (1978), *Theoretical Sensitivity*, Mill Valley, Sociology.
- GLASER, B. G. et A. L. STRAUSS (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Chicago, Aldine.
- HARKER, R. M., D. DOBEL-OBER, D. BERRIDGE et R. SINCLAIR (2004), « More than the sum of its parts? Inter-professional working in the education of looked after children », *Children and Society*, 18(3), 179-193.
- MATTESICH, P., M. MURRAY-CLOSE et B. MONSEY (2001), *Collaboration: What Makes It Work. A Review of Research Literature on Factors Influencing Successful Collaboration*, 2^e éd., Saint Paul, Amerst H. Wilder Foundation.
- PATTON, M. Q. (1990), *Qualitative Evaluation and Research Methods*, 2^e éd., Newbury Park, CA, Sage Publications.
- POSAVAC, E. J. et R. G. CAREY (1992), *Program Evaluation: Methods and Case Studies*, 2^e éd., Englewood Cliffs, Prentice-Hall.
- REITAN, T. C. (1998), « Theories of interorganizational relations in the human services », *Social Service Review*, 72(3), 285-309.

- ROSSI, P. H. (1992), « Assessing family preservation programs », *Children and Youth Services Review*, 14, 77-97.
- ROSSI, P. H. (1992), « Strategies for evaluations », *Children and Youth Services Review*, 14, 167-191.
- SAN MARTIN-RODRIGUEZ, L., M. D. BEAULIEU, D. D'AMOUR et M. FERRADA-VIDELA (2005), « The determinants of successful collaboration: A review of theoretical and empirical studies », *Journal of Interprofessional Care*, 19(s1), 132-147.
- SCOTT, D. (2005). « Inter-organizational collaboration in family-centred practice: A framework for analysis and action », *Australian Social Work*, 58(2), 132-141.
- TANDON, S. D., S. S. AZELTON, J. G. KELLY et D. A. STRICKLAND (1998), « Constructing a tree for community leaders: Contexts and processes in collaborative inquiry », *American Journal of Community Psychology*, 26, 669-696.
- WILDRIDGE, V., S. CHILDS, L. CAWTHRA et B. MADGE (2004), « How to create successful partnerships: A review of the literature », *Health Information and Libraries Journal*, 21(s1), 3-19.
- WORTHINGTON, C. (2005), Patient satisfaction with health care: Recent theoretical developments and implications for evaluation practice, *Revue canadienne d'évaluation de programmes*, 20(3), édition spéciale 2005, 41-63.

Glossaire

Françoise Jabot et Murielle Bauchet

L'évaluation de programmes est une pratique en constante évolution. Comme les chapitres de cet ouvrage le montrent, la pluralité des approches, des modèles de référence et des acteurs, jointe à la multiplicité des modes d'exercice, amène une profusion de termes, qui sont l'objet d'interprétations diverses, voire divergentes. Nous proposons ici un glossaire des termes les plus fréquemment utilisés dans leurs définitions les plus communément acceptées, sans pour autant avoir la prétention de normaliser le champ.

Bénéficiaires : Personnes ou organismes tirant profit de la mise en œuvre d'une intervention, directement ou indirectement, intentionnellement ou non. On distingue différentes catégories de bénéficiaires : ceux qui sont directement visés par une intervention (*bénéficiaires directs*), ceux qui profitent des conséquences d'une intervention sans en être directement les cibles (*bénéficiaires indirects*), ceux qui sont touchés ultimement par l'intervention et qui profiteront de ses conséquences à long terme (*bénéficiaires ultimes*), et ceux qui bénéficient d'un appui dans le cadre de cette intervention (*bénéficiaires intermédiaires*).

Cohérence : Critère de jugement qui interroge l'adéquation entre les différentes composantes de l'intervention : objectifs, ressources allouées, activités prévues (*cohérence interne*) ; ou entre l'intervention évaluée et d'autres interventions visant les mêmes effets (*cohérence externe*).

Coût-avantage : L'analyse coût-avantage mesure l'ensemble des coûts et des avantages sociaux et économiques d'une intervention. Elle implique l'utilisation de prix fictifs qui s'expriment en coût d'opportunité. Le coût d'opportunité se

fonde sur la valeur à laquelle on renonce en optant pour cette intervention plutôt que pour une autre.

Coût-bénéfice : L'analyse coût-bénéfice mesure l'ensemble des coûts et des effets (connexes, directs et indirects, positifs ou négatifs) d'une intervention, échelonnés dans le temps et exprimés en unité monétaire. Elle permet de hiérarchiser différentes options afin de déterminer celle qui présente le meilleur rapport coût-bénéfice.

Coût-efficacité : L'analyse coût-efficacité compare diverses interventions bénéficiant d'un même niveau de ressources (humaines, matérielles, financières). Elle vise à identifier l'intervention la plus satisfaisante pour un coût donné. Les coûts sont exprimés en unité monétaire, et l'effet, en unité naturelle ou physique (*jours d'incapacité évités, années de vie gagnées...*).

Coût-utilité : L'analyse coût-utilité peut porter sur un ou plusieurs effets qui ne sont pas nécessairement communs aux différentes interventions. Elle est basée sur la notion d'utilité, qui s'exprime par la mesure des préférences des individus par rapport à différentes options. Cette analyse est surtout utilisée pour étudier les effets des programmes sur la survie ou la qualité de vie (*années de vie pondérées par la qualité de vie*).

Critère : Le critère est la base du jugement évaluatif et conditionne les types d'évaluation choisis. C'est l'angle de vue adopté pour apprécier une intervention. Cohérence, pertinence, efficacité, efficience et utilité sont les principaux critères qui orientent les questions d'évaluation. Toutefois, pour chaque évaluation, des critères plus spécifiques sont construits et négociés avec l'ensemble des acteurs concernés par l'évaluation, lesquels ne portent pas nécessairement le même regard sur l'intervention. Par exemple, l'évaluation d'un service peut être appréciée sur la base de plusieurs critères, en fonction des intérêts des différentes personnes : l'*accessibilité*, la *permanence du fonctionnement* et la *qualité de l'accueil* pour les usagers de ce service ; la *conformité aux normes techniques* et la *compétence des professionnels* pour les responsables de l'organisme pourvoyeur de services ; l'*adaptabilité du service* aux besoins du fonctionnement pour les professionnels de ce service ; et enfin la *rentabilité* pour le financeur.

Effets (Outcome) : Ce sont les changements observés du fait d'une intervention sur le milieu et les bénéficiaires ; ils peuvent être attendus ou non, attribuables directement ou indirectement à une action, identifiables à court ou moyen terme (*acquisition de connaissances, amélioration des compétences, modification de comportement...*).

Efficacité : Ce critère permet d'apprécier la réalisation des objectifs d'une intervention en comparant les résultats obtenus aux résultats attendus, ces effets étant imputables à cette intervention (*efficacité propre*).

Efficience: L'efficience met en rapport les ressources (budgétaires, humaines, matérielles) mises en œuvre et les effets d'une intervention. Cela suppose d'avoir fait la preuve de l'efficacité. Une intervention efficiente est une intervention efficace au moindre coût.

Évaluabilité: Étude préliminaire (préévaluation) destinée à déterminer si l'évaluation envisagée sera utile et faisable. Elle clarifie la structure de l'intervention, détermine les ressources disponibles, l'accessibilité des données nécessaires, les futurs utilisateurs et leurs besoins ainsi que les méthodes les plus adaptées.

Évaluation: Exercice consistant à répondre à un ensemble de questions relatives à la raison d'être, à la mise en œuvre et aux effets d'une intervention, afin de construire un jugement sur sa valeur à partir de plusieurs angles de vue (efficacité, efficience, pertinence, cohérence...). On recense une grande variété de définitions de l'évaluation. Elle est souvent qualifiée en fonction des objets sur lesquels elle porte (par exemple, évaluation de politiques, de programmes, de projets, de pratiques professionnelles, de réseaux), de ses finalités (évaluation formative, sommative, managériale...), de la discipline principalement mobilisée (évaluation épidémiologique, évaluation économique...), de l'opérateur (évaluation externe, interne, autoévaluation), du moment auquel elle est réalisée (évaluation ex-ante, concomitante ou *in itinere*, ex-post) ou des approches (évaluation participative, collaborative, pluraliste, experte, etc.).

Évaluation concomitante: Intermédiaire, *in itinere* ou chemin faisant, cette évaluation est réalisée au cours de la mise en œuvre d'une intervention. Elle fournit un diagnostic d'étape sur les premières phases de la réalisation. Elle est utile pour vérifier la nécessité de réorienter une intervention.

Évaluation d'impact en santé: Évaluation des effets potentiels d'une intervention (mesure, politique, programme, projet) dans un domaine précis (transport, logement, etc.) sur la santé de la population et évaluation de la distribution de ces effets au sein de la population, en vue d'éclairer les décideurs politiques sur les choix futurs de manière à maximiser les effets potentiels positifs et à minimiser les effets négatifs.

Évaluation ex-ante: Réalisée avant la mise en œuvre d'une intervention, cette évaluation vise à identifier les effets potentiels et la faisabilité d'une intervention.

Évaluation ex-post: Réalisée à la fin d'une intervention, cette évaluation a pour but d'apprécier les effets de cette intervention, d'en identifier les facteurs de réussite ou d'échec, et d'en tirer des enseignements aux fins d'autres interventions.

Évaluation externe: L'évaluation externe est conduite par des personnes non impliquées directement dans la mise en œuvre de l'intervention évaluée.

Évaluation formative: Cette évaluation est effectuée en vue d'améliorer une intervention en cours de mise en œuvre. Elle vise à éclairer les acteurs de l'intervention sur le contexte et les conséquences de leurs pratiques.

Évaluation interne: L'évaluation interne (ou autoévaluation) est effectuée par les acteurs qui prennent part activement à une intervention. Elle est utile pour entretenir de manière continue un questionnement vis-à-vis du travail effectué et pour réajuster les programmes et actions en cours. Cependant, la distinction entre évaluation interne et évaluation externe peut être difficile à faire dans la mesure où l'activité évaluative peut adopter des formes mixtes: lorsqu'elle est conduite par l'organisation responsable de l'intervention à évaluer et supervisée par un acteur extérieur (autoévaluation accompagnée) ou lorsqu'un service d'une institution évalue l'activité d'un autre service de cette même institution.

Évaluation participative: L'évaluation participative accorde une place particulière à toutes les personnes concernées (y compris les bénéficiaires ultimes) par une intervention et son évaluation. Elle prend en considération leurs valeurs, leurs points de vue, leurs intérêts et leurs attentes, et cela à toutes les étapes du processus. Elle peut cependant prendre des formes diverses en fonction de la nature et de la profondeur de la participation recherchée: l'évaluation pratique, axée sur la participation des parties prenantes à la prise de décision; l'évaluation émancipatrice, qui cherche à renforcer leur capacité et leur pouvoir d'agir (*empowerment*).

Évaluation sommative (ou récapitulative): Ce type d'évaluation, effectué au terme d'une intervention, a pour but de donner une appréciation globale et distanciée principalement sur les effets d'une intervention; elle est habituellement utilisée pour orienter la prise de décision quant au devenir de cette dernière (étendre à d'autres clientèles, à d'autres territoires, diminuer ou augmenter les ressources, modifier ou supprimer l'intervention, etc.).

Extrants (réalisations, *outputs*): Ce sont les productions de l'intervention: produits, biens et services (par exemple, *sessions de formation, consultations, conférences, débats, rencontres de parents, documents pédagogiques*).

Faisabilité d'un programme: L'étude de faisabilité, effectuée lors de la conception d'une intervention, vise à s'assurer que celle-ci est réalisable sur les plans économique, technique et organisationnel.

Impact: Ensemble des effets sociaux, économiques et environnementaux à long terme imputables à une intervention, qu'ils soient positifs ou négatifs, souhaités ou non, immédiats ou différés. Par exemple, l'amélioration de la qualité de vie, la réduction de la mortalité. Cependant, il est très difficile d'attribuer les impacts observés à une intervention précise, étant donné les interactions possibles avec d'autres interventions et les multiples facteurs susceptibles d'entrer en jeu.

Indicateur: Information permettant d'objectiver une réalité et n'ayant de sens que confrontée à un critère. L'indicateur peut être quantitatif ou qualitatif. On en distingue différents types: indicateurs de contexte, qui qualifient une situation; indicateurs de réalisation, qui décrivent ce qui a été fait (*nombre de per-*

sonnes dépistées, nombre de jours de consultation...); indicateurs de performance, qui renseignent sur l'atteinte des objectifs (pourcentage de personnes dépistées, délai d'attente de consultation); indicateurs d'impact, qui attestent des effets de l'intervention (nombre de décès évités, changements de comportement, adoption de pratiques nouvelles...).

Intervention : Ce terme générique désigne une action ou un ensemble organisé d'actions (politique, programme, projet) visant à modifier une situation jugée problématique.

Intrants (Inputs) : Ressources matérielles, humaines et financières allouées pour la mise en œuvre d'une intervention.

Modèle logique : Outil permettant de formaliser la logique d'une intervention, à savoir les relations entre les différentes composantes de cette intervention (buts, objectifs, activités, effets attendus), de visualiser les différentes voies possibles pour atteindre les résultats (les chaînes de causalité) et d'identifier les facteurs externes susceptibles d'influencer cette intervention.

Pérennisation : Processus permettant la poursuite d'une intervention, une fois l'assistance financière et technique terminée. Il s'agit d'un processus de routinisation qui aboutit à la pérennité d'une intervention.

Pérennité : Une intervention est pérenne quand elle est capable de produire des bénéfices pour la population visée qui perdurent une fois l'intervention terminée. Le degré de pérennité porte sur des caractéristiques routinières d'une organisation : maintien et stabilité des ressources (fonds dédiés, temps dévolu...), adaptation de l'intervention au contexte et à ses effets, valeurs associées et règles d'intervention partagées avec l'organisation (partage des objectifs, nomination d'un responsable, existence de procédures...).

Performance : Capacité de produire des résultats conformes aux objectifs fixés et à un coût acceptable.

Pertinence : Critère interrogeant le bien-fondé d'une intervention, à savoir l'adéquation entre ses objectifs et la nature du problème à résoudre.

Référentiel : Cadre d'interprétation des données recueillies aux fins de l'évaluation et à partir duquel un jugement peut être formulé. C'est un système de référence qui permet de confronter les indicateurs (éléments représentatifs d'une situation réelle) aux critères et aux normes ou valeurs de référence, quand celles-ci existent (éléments significatifs d'une situation désirée) dans un contexte donné.

Suivi (monitorage) : Processus continu de collecte de données selon des indicateurs préalablement définis et visant à fournir aux responsables d'une intervention des informations sur les réalisations en cours, l'atteinte des objectifs et l'utilisation des ressources allouées.

Tableau de bord : Outil d'aide à la gestion regroupant un ensemble d'indicateurs spécifiques recueillis régulièrement pour le suivi d'une intervention.

Théorie d'action (théorie de programme) : Stratégie sur laquelle est construite une intervention, notamment les hypothèses émises pour expliquer la façon dont celle-ci devrait produire ses effets et atteindre son objectif global. La théorie de programme est centrée sur les hypothèses alors que le modèle logique est la représentation graphique des activités du programme et des effets attendus.

Utilisation de l'évaluation : L'utilisation de l'évaluation peut revêtir diverses formes selon les utilisateurs (acteurs de l'intervention évaluée, planificateurs, décideurs...), les buts avoués ou implicites (éclairer la décision, améliorer l'action, résoudre un problème, justifier ou soutenir une décision, etc.). Une évaluation commanditée pour éclairer la prise de décision publique sera préférentiellement utilisée sur la base de la connaissance apportée par les travaux d'évaluation et accessible dans le rapport d'évaluation (*utilisation centrée sur les résultats*). Cependant, le processus même d'évaluation d'une intervention, en raison des interactions entre les différents acteurs concernés, de leurs occasions d'action et de leurs apprentissages, est susceptible de produire des changements qui affecteront cette intervention et le contexte dans lequel elle est mise en œuvre (*utilisation centrée sur le processus*). L'utilisation d'une évaluation dépend de plusieurs paramètres, notamment des caractéristiques des utilisateurs, de la conduite de l'évaluation et du contexte dans lequel elle se déroule. L'expertise et les habiletés de l'évaluateur dans l'animation du processus jouent également un rôle très important.

Sources

BANQUE DE DONNÉES EN SANTÉ PUBLIQUE (BDSP), *Glossaire multilingue*, www.bdsp.tm.fr/Glossaire/

BUREAU, D. et M. MOUGEOT (2007), *Performance, incitations et gestion publique*, Paris, La Documentation française.

COLLÈGE DES ÉCONOMISTES DE LA SANTÉ (2003), *Guide méthodologique pour l'évaluation économique des stratégies de santé*: www.ces-asso.org/Pages/default_fr.htm

COMMISSION EUROPÉENNE (2001), *Gestion du cycle de projet*, Unité d'évaluation de l'office de coopération EuropAid.

CONSEIL SCIENTIFIQUE DE L'ÉVALUATION (1996), *Petit guide d'évaluation des politiques publiques*, Paris, La Documentation française.

DRUMMOND, MICHAEL F. et al. (1997), *Méthodes d'évaluation économique des programmes de santé*, 2^e éd., Paris, Economica.

FEDER (2003), *Guide de l'analyse coûts-avantages des grands projets*.

- LEVITON, L. C. *et al.* (2010), « Evaluability assessment to improve public health policies, programs, and practices », *Annual Review of Public Health*, 31, 213-233.
- MATHISON, S. (2005), *Encyclopedia of Evaluation*, Thousand Oaks/Londres, Sage Publications.
- ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2002), *Glossaire des principaux termes relatifs à l'évaluation et à la gestion axée sur les résultats*, Bruxelles, OCDE.
- ORGANISATION MONDIALE DE LA SANTÉ (1999), *Göteborg Consensus Paper*.
- PATTON, M. Q. (1997), *Utilization-Focused Evaluation*, Thousand Oaks, Sage Publications.
- PERRET, B. (2001), *L'évaluation des politiques publiques*, Paris, La Découverte.
- RIDDE, V. (2006), « Suggestions d'améliorations d'un cadre conceptuel de l'évaluation participative », *The Canadian Journal of Program Evaluation*, 21(2), 1-23.
- RIDDE, V. et P. PLUYE (2006), « Évaluer la pérennité des programmes de santé publique : un outil et son application en Haïti », *Revue d'épidémiologie et de santé publique*, 54, 421-431.
- SCRIVEN, M. (1991), *Evaluation Thesaurus*, 4^e éd., Newbury Park, CA, Sage Publications.
- UNION EUROPÉENNE, *Evaluation of the socio-economic development. Resources for evaluation, the guide* : www.evaled.info

Sigles et abréviations

ASACO	Association de santé communautaire
ATTRueQ	Association des travailleurs et travailleuses de rue du Québec
CEQ	Centrale de l'enseignement du Québec
CIRCM	Centre international de résolution de conflits et de médiation
CJM-IU	Centre jeunesse de Montréal – Institut universitaire
CNPC	Centre national de prévention du crime du Canada
CPS	Centre de prévention du suicide
CRDI	Centre de recherches pour le développement international
CRISE	Centre de recherche et d'intervention sur le suicide et l'euthanasie
CSCOM	Centre de santé communautaire
CSRéf	Centre de santé de référence
DDSES	Direction du développement social et de l'économie solidaire
DNS	Direction nationale de la santé
DPJ	Directeur de la protection de la jeunesse
DRS	Direction régionale de la santé
ERPI	Entrevue de résolution de problèmes interpersonnels
GEMMS	Groupe d'étude sur la médiation en milieu scolaire
IRI-Accueil	Intervention rapide et intensive rattachée à l'accueil
IRSST	Institut de recherche Robert-Sauvé en santé et en sécurité du travail
MM	Méthodes mixtes
PED	Pays en développement
PIJ	Programme intégration jeunesse
QES	Questionnaire sur l'environnement socioéducatif
QL ou QUAL	Méthodes qualitatives
QT ou QUANT	Méthodes quantitatives
RAJB	Ressource alternative des jeunes de Bellechasse
RLM	Revue de littérature mixte (revue des études QL, QT et MM)
RTS	Réception-traitement des signalements
SAM	Suicide Action Montréal

SOU	Soins obstétricaux d'urgence
SRE	Système de référence-évacuation
SST	Santé et sécurité du travail
TR	Travailleurs de rue
UdeM	Université de Montréal
UNFPA	Fonds des Nations Unies pour la population
US	Urgences sociales

Bibliographie

- ABBOTT, A. (1998), « The causal devolution », *Sociological Methods and Research*, 27, 148-181.
- ABMA, T. A. (1997), « Playing with/in plurality: Revitalizing realities and relationships in Rotterdam », *Evaluation*, 3(1), 25-48.
- ADLER, P. S. (1999), « Building better bureaucracies », *Academy of Management Executive*, 13(4), 36-47.
- ADLER, P. S. et B. BORYS (1996), « Two types of bureaucracy: Enabling vs coercive », *Administrative Science Quarterly*, 41(1), 61-89.
- AGUILERA, D. C. (1990), *Crisis Intervention: Theory and Methodology*, St-Louis, C. V. Mosby.
- ALKIN, M. (dir.) (1990), *Debates on Evaluation*, Newbury Park, CA, Sage Publications.
- ALKIN, M. et K. COYLE (1988), « Thoughts on evaluation misutilization », *Studies in Educational Evaluation*, 14, 331-340.
- ALVESSON, M. (1993), « Organization as rhetoric: Knowledge-intensive firms and the struggle with ambiguity », *Journal of Management Studies*, 30(6), 997-1015.
- ALVESSON, M. (2001), « Knowledge work: Ambiguity, image and identity », *Human Relations*, 54(7), 863-886.
- ARCHAMBAULT, J. et C. DES GROSEILLERS (1997), *Intervention rapide et intensive en amont en situation d'urgence et de crise*, Les Centres jeunesse de Montréal, Direction des services territoriaux.
- ASHKENAZI, M. (1986), *Research Methods in Social Sciences (Unit 4: Ethnographic Research in Anthropology)*, Tel Aviv, Open University (en hébreu).
- AUBEL, J. (2000), *Manuel d'évaluation participative de programme*, 2^e éd., Calverton, CSTS-CRS-USAID.

- AUERBACH, S. M. et A. L. STOLBERG (1986), *Crisis Intervention with Children and Families*, Washington, Hemisphere.
- BACHRACH, P. et M. S. BARATZ (1963), «Decisions and nondecisions: An analytical framework», *American Political Science Review*, 57(3), 632-642.
- BAIZERMAN, M., D. W. COMPTON et S. H. STOCKDILL (2004), «Capacity Building», dans S. Mathison (dir.), *Encyclopedia of Evaluation*, Thousand Oaks, Sage Publications.
- BAMBERGER, M. (2006), *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints*, Washington, Banque mondiale.
- BAMBERGER, M., J. RUGH et L. MABRY (2006), *RealWorld Evaluation: Working under Budget, Time, Data and Political Constraints*, Thousand Oaks, Sage Publications.
- BANQUE MONDIALE (2004), *Influential Evaluations: Evaluations That Improved Performance and Impacts of Development Programs*, Washington, Banque mondiale, Operations Evaluation Department.
- BANQUE MONDIALE (2005), *Influential Evaluations: Detailed Case Studies*, Washington, Banque mondiale, Operations Evaluation Department.
- BARDACH, E. (1977), *Implementation Game*, Cambridge, MIT Press.
- BARNEY, J. (1991), «Firm resources and sustained competitive advantage», *Academy of Management Executive*, 17(1), 99-120.
- BAUMGARTNER, F. et B. JONES (1993), *Agendas and Instability in American Politics*, University of Chicago Press.
- BEAUDRY, J. et B. GAUTHIER (1992), «L'évaluation de programmes», dans B. Gauthier (dir.), *Recherche sociale: de la problématique à la collecte de données*, 2^e éd., Sainte-Foy, Presses de l'Université du Québec, 425-452.
- BEAULÉ, G. et P. SIMARD (2002), *L'approche de proximité en milieu rural: quel modèle pour le Témiscamingue?*, Régie régionale de la santé et des services sociaux de l'Abitibi-Témiscamingue.
- BECHLER, Z. (1987), *Philosophy of Science*, Tel Aviv, University on the Air, Israel Ministry of Defense (en hébreu).
- BÉGIN, C., P. JOUBERT et J. TURGEON (1999), «L'évaluation dans le domaine de la santé: conceptions, courants de pensée et mise en œuvre», dans C. BÉGIN et al. (dir.), *Le système de santé québécois: un modèle en transformation*, Montréal, Presses de l'Université de Montréal, 265-281.
- BENSON, J. K. (1975), «The interorganizational network as a political economy», *Administrative Science Quarterly*, 2(2), 229.
- BERTHELETTE, D. et al. (1998), «Évaluation de l'implantation d'un programme de formation en santé et en sécurité du travail», *Performances humaines et techniques*, 21-26.

- BERTHELETTE, D., L. DESNOYERS et A. BÉDARD (2001), « Recherche évaluative d'un programme syndical de formation en santé et en sécurité du travail », *Relations industrielles*, 56(3), 516-542.
- BILODEAU, A. et al. (2006), « Les dispositifs de la participation aux étapes stratégiques de l'évaluation », *Revue canadienne d'évaluation de programmes*, 21(3), 257-282.
- BITONI, C. (1993), « Cognitive mapping: A qualitative research method for social work », *Social Work Research and Abstracts*, 29(1), 9-16.
- BLAISE, P., B. MARCHAL, P. LEFÈVRE et G. KEGELS (2010), « Au-delà des méthodes expérimentales, l'approche réaliste en évaluation », dans L. POTVIN, M. J. MOQUET et C. JONES (dir.), *Réduire les inégalités sociales de santé*, Saint-Denis, INPES, coll. « Santé en action ».
- BLONDIN, M. (1980), « Une formation syndicale faite par les travailleurs eux-mêmes », *Revue internationale d'action communautaire*, 3(43), 73-80.
- BLOOM, B. S. et al. (1969), *Taxonomie des objectifs pédagogiques*, t. 1: *Domaine cognitif*, Montréal, Les Entreprises éducation nouvelle.
- BOURDIEU, P. et L. WACQUANT (1992), *Réponses: pour une anthropologie réflexive*, Paris, Éditions du Seuil.
- BOWEN, F. et al. (2003), *Analyse des impacts d'un programme de promotion des conduites pacifiques du CIRCM: Rapport préliminaire pour la première année de mise en œuvre (2001-2002)*, Ottawa, rapport présenté au Centre national de prévention du crime.
- BOWEN, F. et al. (2006), *Rapport final d'évaluation des impacts du programme « Vers le pacifique » pour les quatre années de mise en œuvre (2001-2005)*, Ottawa, rapport présenté au Centre national de prévention du crime.
- BRENT, D. A. (1995), « Risk factors for adolescent suicide and suicidal behavior: Mental and substance abuse disorders, family environmental factors and life stress », *Suicide and Life-Threatening Behavior*, 25 (suppl.), 52-63.
- BREWER, G. D. et Peter DE LEON (1983), *The Foundations of Policy Analysis*, Pittsburgh, University of Pittsburgh Press.
- BRISOLARA, S. (1998), « The history of participatory evaluation and current debates in the field », dans E. WHITMORE (dir.), *Understanding and Practicing Participatory Evaluation: New Directions for Evaluation*, n° 80, Hoboken, Jossey-Bass, 25-41.
- BROWN, J. et P. DUGUID (1991), « Organizational learning and communities of practice: Toward a unified view of working, learning and innovation », *Organization Sciences*, 2(1), 40-57.
- BRYMAN, A. (2006), « Integrating quantitative and qualitative research: How is it done? », *Qualitative Research*, 6(1), 97-113.
- BUSSMANN, W., U. KLÖTI et P. KNOEPFEL (dir.) (1998), *Politiques publiques: évaluation*, Paris, Économica.

- CAMPBELL, D. T. (1957), « Factors relevant to the validity of experiments in social settings », *Psychological Bulletin*, 54, 297-312.
- CAMPBELL, D. T. (1986), « Relabeling internal and external validity for applied social sciences », *New Direction for Program Evaluation*, 31, 67-77.
- CAMPBELL, D. T. (1988), « Qualitative knowing in action research », dans S. OVERMAN (dir.), *Methodology and Epistemology for Social Science: Selected Papers of Donald T. Campbell*, Chicago, University of Chicago Press, 360-376.
- CAMPBELL, D. T. et D. W. FISKE (1959), « Convergent and discriminant validation by the multitrait-multimethod matrix », *Psychological Bulletin*, 56, 81-105.
- CAMPBELL, D. T. et J. C. STANLEY (1966), *Experimental and Quasi-experimental Design for Research*, Boston, Houghton-Mifflin.
- CAMPBELL, O. M. et W. J. GRAHAM (1990), *Measuring Maternal Mortality and Morbidity: Level and Trends*, Maternal and Child Epidemiology Unit Publication No. 2, London School of Hygiene and Tropical Medicine.
- CAPLAN, G. (1964), *Principles of Preventive Psychiatry*, New York, Basic Books.
- CAPLAN, N., A. MORRISON et R. J. STAMHAUGH (1975), *The Use of Social Science Knowledge in Public Policy Decisions at the National Level*, Ann Arbor, Institute for Social Research.
- CAPP, K., F. P. DEANE et G. LAMBERT (2001), « Suicide prevention in aboriginal communities: Application of community gatekeeper training », *Australian and New Zealand Journal of Public Health*, 25(4), 315-321.
- CARROLL, C. et al. (2007), « A conceptual framework for implementation fidelity », *Implement Science*, 2(40).
- CAWS, P. (1965), *The Philosophy of Science*, Princeton, D. Van Nostrand.
- CHAMPAGNE, F., L. LEMIEUX-CHARLES et W. MCGUIRE (2004), « Introduction: Towards a broader understanding of the use of knowledge and evidence », dans L. LEMIEUX-CHARLES et F. CHAMPAGNE (dir.), *Using Knowledge and Evidence in Health Care. Multidisciplinary Perspective*, Toronto, University of Toronto Press.
- CHAMPAGNE, F. et J.-L. DENIS (1990), « Pour une évaluation sensible à l'environnement des interventions: analyse de l'implantation », *Service social*, 41(1), 143-193.
- CHEN, H. (1989), « The theory-driven perspective », *Evaluation and Program Planning*, 12, 4.
- CHEN, H. (1990), *Theory-Driven Evaluation*, Newbury Park, CA, Sage Publications.
- CHEN, H. (2005), *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness*, Thousand Oaks, Sage Publications.
- CHEN, H. (2005), « Theory-Driven Evaluation », dans S. MATHISON (dir.), *Encyclopedia of Evaluation*, Thousand Oaks, Sage Publications, 415-419.

- CHEN, H. et P. H. ROSSI (1981), « The multi-goal theory-driven approach to evaluation : A model linking basic and applied social sciences », dans H. Freeman et M. SOLOMAN (dir.), *Evaluation Studies Review Annual*, n° 6, Newbury Park, CA, Sage Publications, 38-54.
- CHEN, H. et P. H. ROSSI (1983), « Evaluating with sense: The theory-driven approach », *Evaluation Review*, 7(3), 283-302.
- CHEN, H. et P. H. ROSSI (1992), « Introduction : Integrating theory into evaluation practice », dans H. CHEN et P. H. ROSSI (dir.), *Using Theory to Improve Program and Policy Evaluation*, Westport, Greenwood Publishing Group, 1-11.
- CHENITZ, W. et J. SWANSON (1986), *From Practice to Grounded Theory: Qualitative Research in Nursing*, Menlo Park, Addison-Wesley.
- CHRISTENSEN, C. M., R. BOHMER et J. KENAGY (2000), « Will disruptive innovations cure health care? », *Harvard Business Review*, sept.-oct., 102-112.
- COBB, R. W. et D. A. ROCHEFORT (1993), « Problem definition, agenda access, and policy coice », *Policy Studies Journal*, 1(1), 56-71.
- COBB, R. W. et M. H. ROSS (1997), « Agenda setting and the denial of agenda access: Ky concepts », dans R. W. COBB et M. H. ROSS (dir.), *Cultural Strategies of Agenda Denial: Avoidance, Attack and Redefinition*, Lawrence, University Press of Kansas.
- COHEN, W. M. et D. A. LEVINTHAL (1990), « Absorptive capacity: A new perspective on learning and innovation », *Administrative Science Quarterly*, 35(1).
- COMPTON, D. W. et al. (2001), « Developing evaluation capacity while improving evaluation training in public health: The American Cancer Society's Collaborative Evaluation Fellows Project », *Evaluation and Program Planning*, 24(1), 33-40.
- COOK, T. D. (1985), « Post-positivist critical multiplism », dans L. SHOTLAND et M. MARK (dir.), *Social Science and Social Policy*, Beverly Hills, Sage Publications.
- COOK, T. et D. T. CAMPBELL (1976), « The design and conduct of quasi-experiments and true experiments in field settings », dans M. DUNNETTE (dir.), *Handbook of Industrial and Organizational Psychology*, Chicago, Rand McNally.
- COOK, T. et D. T. CAMPBELL (1979), *Quasi-experiments: Design and Analysis for Field Settings*, Chicago, Rand McNally.
- COPI, I. M. (1961), *Introduction to Logic*, Londres, Macmillan.
- COPI, I. M. et K. BURGESS-JACKSON (1995), *Informal Logic*, Up Saddle River, NJ, Prentice Hall.
- CORDRAY, D. S. (1986), « Quasi-experimental analysis: A mixture of methods and judgment », *New Directions for Program Evaluation*, 31, 9-28.
- COUSINS, J. B. (2004), « Commentary : Minimizing evaluation misuse as principled practice », *American Journal of Evaluation*, 25(3), 391-397.

- COUSINS, J. B. (2005), « Will the real empowerment evaluation please stand up ? A critical friend perspective », dans D. M. FETTERMAN et A. WANDERSMAN (dir.), *Empowerment Evaluation Principles in Practice*, New York, Guilford Press, 183-208.
- COUSINS, J. B. (2006), « Rôle du gouvernement dans l'assurance de la qualité de l'évaluation », document de travail préparé pour le Centre d'excellence en évaluation du Secrétariat du Conseil du Trésor du Canada, Ottawa. En ligne : www.tbs-sct.gc.ca/eval/tools_outils/QAE-AQE/rgeqa_f.asp
- COUSINS, J. B. et E. WHITMORE (1998), « Framing participatory evaluation », dans E. WHITMORE (dir.), *Understanding and Practicing Participatory Evaluation : New Directions for Evaluation*, n° 80, San Francisco, Jossey-Bass, 5-42.
- COUSINS, J. B. et L. M. SHULHA (2006), « A comparative analysis of evaluation utilization and its cognate fields of Inquiry », dans I. SHAW, J. C. GREENE et M. M. MARK (dir.), *The Sage Handbook of Evaluation*, Thousand Oaks, Sage Publications.
- COUSINS, J. B., J. DONOHUE et J. BLOOM (1996), « Collaborative evaluation in North America: Evaluators' self-reported opinions, practices and consequences », *Evaluation Practice*, 17(3), 207-226.
- CPS/MS, DNSI et ORC Macro (2007), *Enquête démographique et de santé au Mali 2006*, Cellule de planification et de statistique du ministère de la Santé (CPS/MS), Direction nationale de la statistique et de l'informatique (DNSI) et ORC Macro International, www.measuredhs.com
- CRESWELL, J. W. (1998), *Qualitative Inquiry and Research Design: Choosing among Five Traditions*, Thousand Oaks, Sage Publications.
- CRESWELL, J. W. et al. (2003), « Teaching mixed methods research », dans A. TASHAKKORI et C. TEDDLIE (dir.), *Handbook of Mixed Methods in Social and Behavioral Research*, Thousand Oaks, Sage Publications, 619-637.
- CRESWELL, J. W. et V. PLANO CLARK (2007), *Designing and Conducting Mixed Methods Research*, Thousand Oaks, Sage Publications.
- CRESWELL, J. W., M. D. FETTERS et N. V. IVANKOVA (2004), « Designing a mixed methods study in primary care », *Annals of Family Medicine*, 2(1), 7-12.
- CRONBACH, L. J. (1963), « Course improvement through evaluation », *Teachers College Record*, n° 64, 672-683.
- CRONBACH, L. J. (1983), *Designing Evaluation of Educational and Social Programs*, San Francisco, Jossey-Bass.
- CROZIER, M. et E. FRIEDBERG (1977), *L'acteur et le système*, Paris, Éditions du Seuil
- CUMMINGS, J. (2003), *Knowledge Sharing: A Review of the Literature*, Washington, World Bank Operations Evaluation Department.
- DAGENAIS, C. et C. BOUCHARD (1993), « Les programmes de soutien intensif aux familles. Intervention massive ou intervention magique ? », *PRISME*, 3, 503-515.

- DAGENAIS, C. et C. BOUCHARD (1996), « L'impact des programmes de soutien intensif visant à maintenir les enfants et adolescents dans leur famille », *Revue canadienne de santé mentale communautaire*, 15(1), 63-82.
- DAGENAIS, C., J. BÉGIN et D. FORTIN (1999), « L'évaluation des services intensifs offerts à des familles en crise et l'amélioration du fonctionnement familial. Une méta-analyse », *Les cahiers d'analyse du GRAVE*, 6(1), Université du Québec à Montréal, Laboratoire de recherche en écologie humaine et sociale.
- DAGENAIS, C. *et al.* (2003), « Évaluation de l'implantation et des effets d'un programme de soutien intensif offert à des familles afin d'éviter un placement en milieu substitut », *Revue canadienne d'évaluation de programmes*, 18(2), 47-69.
- DAGENAIS, C. *et al.* (2004), « Impact of intensive family support programs: A synthesis of evaluation studies », *Children and Youth Services Review*, 26, 249-263.
- DAGENAIS, C. (2010), « Knowledge transfer in community based organizations (CBOs): A needs assessment study », *Global Journal of Community Psychology Practice*, 1(2), 13-30. Retrieved 2010 09 22 from www.gjcpp.org/
- DAGENAIS, C. *et al.* (sous presse). Use of evidence by school practitioners and determinants of use: A review of empirical research. Evidence and Policy.
- DAHL, R. A. (1961), *Who Governs?*, Yale University Press.
- DAHLGREN, G. et M. WHITEHEAD (1992), *Policies and Strategies to Promote Equity in Health*, Copenhagen, WHO Regional Office for Europe.
- DAVID, G. (1987), « Nécessité et efficacité d'une approche intégrée dans le traitement de l'inceste », *Service social*, 36, 350-368.
- DAVIDSON, E. J. (2001), « Mainstreaming evaluation into an organization's learning culture », communication présentée à la Conférence annuelle de l'American Evaluation Association, St Louis, <http://unix.cc.wmich.edu/~j davidso/olcult.ppt>
- DAVIDSON, M. W. et L. M. RANGE (1999), « Are teachers of children and young adolescents responsive to suicide prevention training modules? », *Death Studies*, 23(1), 61-71.
- DAVIES, I. (1999), « Evaluation and performance management in government », *Evaluation*, 5(2), 150-159.
- DAVIES, P. (2004), « Is evidence-based government possible? », communication au 4th Annual Campbell Collaboration Colloquium, Washington.
- DE BERNIS, L. (2003), « La mortalité maternelle dans les pays en développement : quelles stratégies adopter? », *Médecine tropicale*, vol. 63, n^{os} 4-5, 391-399.
- DE LEON, P. (1987), « Policy termination as a political phenomenon », dans D. J. PALUMBO (dir.), *The Politics of Program Evaluation*, Newbury Park, CA, Sage Publications.
- DENIS, J. L. (2004), « Governance and management of change in Canada's health system », dans P. G. FOREST, G. P. MARCHILDON et T. MCINTOSH (dir.), *Romanow Papers*, vol. 2, University of Toronto Press.

- DENIS, J. L., J. LOMAS et N. STIPICH (2008), « Creating receptor capacity for research in the health care system : The executive training for research application (EXTRA) program in Canada » (éditorial), *Journal of Health Services Research and Policy*, 13(1), 1-7.
- DENIS, J.-L., P. LEHOUX et F. CHAMPAGNE (2004), « Knowledge utilization on fine-tuning dissemination and contextualizing knowledge », dans L. LEMIEUX-CHARLES et F. CHAMPAGNE (dir.), *Using Knowledge and Evidence in Health Care. Multidisciplinary Perspectives*, Toronto, University of Toronto Press, 41-69.
- DENIS, J.-L. et al. (2002), « Explaining diffusion patterns for complex health care innovations », *Health Care Management Review*, 27(3), 60-73.
- DENZIN, N. K. et Y. S. LINCOLN (1994), *Handbook of Qualitative Research*, Newbury Park, CA, Sage Publications.
- DESNOYERS, L. (1986), « VDT's and occupational health : Workers education », *Proceedings of the International Conference Work with Display Units*, Stockholm, 108-110.
- DONALDSON, L. (2001), « Reflections on knowledge and knowledge-intensive firms », *Human Relations*, 54(7), 955.
- DORÉ, G. et R. MARCEAU (2006), « L'évaluation de programmes à la fonction publique québécoise : un profil de compétences requises », *Télescope*, printemps-été, 19-30.
- DREWETT, A. (1997), « Evaluation and consultation : Learning the lessons of user involvement », *Evaluation. The International Journal of Theory, Research and Practice*, 3(2), 189-204.
- DRS et DDES-Kayes (2007), *Rapport d'activités 2006. Santé et développement social et Plan opérationnel 2008 – Compilé régional, Région de Kayes*, Direction régionale de la santé et Direction du développement social et de l'économie solidaire de Kayes.
- DRUMMOND, M. F. et al. (1998), *Méthodes d'évaluation économique des programmes de santé*, 2^e éd., Paris, Économica.
- DRYZEK, J. S. (1990), *Discursive Democracy: Politics, Policy and Political Science*, Cambridge University Press.
- DUBOIS, N. (1996), « Le locus of control », dans J.-C. DESCHAMPS et J.-L. BEAUVOIS (dir.), *La psychologie sociale*, t. II : *Des attitudes aux attributions. Sur la construction de la réalité sociale*, Grenoble, Presses universitaires de Grenoble, 227-236.
- DUMONT, A. et al. (2006), *Enquête de faisabilité des audits de décès maternels au Sénégal : rapport final*, Montréal, CHU Sainte-Justine.
- DU RANQUET, M. (1991), *Les approches en service social*, Québec, Edisem.
- DURAN, P. (1996), « Les non-dits de l'évaluation », dans G. TIMSIT, A. CLAISSE et N. BELLOUBET-FRIER (dir.), *Les administrations qui changent. Innovations techniques ou nouvelles logiques?*, Paris, PUF, 161-186.

- DUVAL, M. et A. FONTAINE (2000), « Lorsque des pratiques différentes se heurtent : les relations entre les travailleurs de rue et les autres intervenants », *Nouvelles Pratiques sociales*, 13(1), 49-67.
- EARL, S., F. CARDEN et T. SMUTYLO (2002), *La cartographie des incidences : intégrer l'apprentissage et la réflexion dans les programmes de développement*, Ottawa, CRDI.
- ERIKSON, E. H. (1959), « Identity and the life cycle », *Psychological Issues*, 1, 1-171.
- ESTRALLA, M. (dir.) (2004), *L'évaluation et le suivi participatifs : apprendre du changement*, Paris, CRDI et Karthala.
- EYKEN, W. V., D. GOULDEN et M. CROSSLEY (1995), « Evaluating educational reform in a small state : A case study of Belize, Central America », *Evaluation*, 1(1), 33-44.
- FAFARD, A., A. BOURASSA et S. VEILLEUX (2000), *Portrait Jeunesse de la MRC de Bellechasse*, Régie régionale de la santé et des services sociaux de Chaudière-Appalaches.
- FETTERMAN, D. M. (1996), « Empowerment evaluation : An introduction to theory and practice », dans D. M. FETTERMAN, S. J. KAFFARIAN et A. WANDERSMAN (dir.), *Empowerment Evaluation Knowledge and Tools for Self-Assessment et Accountability*, Thousand Oaks, Sage Publications.
- FETTERMAN, D. M. (2000), *Foundations of Empowerment Evaluation*, Thousand Oaks, CA, Sage Publications.
- FETTERMAN, D. M. (2001), *Foundations of Empowerment Evaluation*, Thousand Oaks, CA, Sage Publications.
- FETTERMAN, D. M. et A. WANDERSMAN (2005), *Empowerment Evaluation Principles in Practice*, New York, Guilford Press.
- FINNE, H., M. LEVIN et T. NILSSEN (1995), « Trailing research : A model for useful program evaluation », *Evaluation*, 1(1), 11-32.
- FLAY, B. R. et al. (2005), « Standards of evidence : Criteria for efficacy, effectiveness and dissemination », *Prevention Science*, 6(3), 151-175.
- FOUREZ, G. (2003), *Apprivoiser l'épistémologie*, Bruxelles, De Boeck.
- FOURNIER, P. et O. OUATTARA (2005), *Suivi-évaluation du système de référence-évacuation et des caisses de solidarité pour les urgences obstétricales de la Région de Kayes*, Rapports de mission de juin 2004 et mars 2005, Montréal, Unité de Santé internationale/Université de Montréal.
- FOURNIER, P. et al. (2006), *Système de référence-évacuation des urgences obstétricales de la Région de Kayes. Rapport de mission*, Montréal, Unité de Santé internationale/Université de Montréal.
- FRAMBACH, R. T. et N. SCHILLEWAERT (2002), « Organizational innovation adoption : A multi-level framework of determinants and opportunities for future research », *Journal of Business Research*, 55, 163-176.

- FREIRE, P. (1983), *Pédagogie des opprimés*, Paris, La Découverte-Maspero.
- GARANT, L. (1992), *Les programmes de soutien familial. Une alternative au placement des jeunes?*, Québec, Ministère de la Santé et des Services sociaux.
- GARRAUD, P. (1990), « Politiques nationales: élaboration de l'agenda », *L'Année sociologique*, 40, 17-41.
- GEERTZ, C. (1973), *The Interpretation of Cultures*, New York, Basic Books.
- GENDRON, S. (2001), *La pratique participative en santé publique: l'émergence d'un paradigme*, Montréal, Université de Montréal.
- GIBBS, D. et al. (2002), « Increasing evaluation capacity within community-based HIV prevention programs », *Evaluation and Program Planning*, 25(3), 261-269.
- GIORDONO, Y. (2003), *Conduire un projet de recherche: une perspective qualitative*, Paris, EMS.
- GLASER, B. et A. L. STRAUSS (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Chicago, Adline.
- GLASER, B. G. (1978), *Theoretical Sensitivity*, Mill Valley, Sociology Press.
- GOETZ, J. P. et M. D. Lecompte (1984), *Ethnography and Qualitative Design in Educational Research*, Londres, Academic Press.
- GOGGIN, M. L. et al. (1990), *Implementation Theory and Practice: Toward a Third Generation*, Glenview, Scott Foresman/Little, Brown.
- GOLAN, N. (1978), *Treatment in Crisis Situations*, New York, Free Press.
- GOULD, M. S. et R. A. KRAMER (2001), « Youth suicide prevention », *Suicide and Life-Threatening Behavior*, 31, 6-30.
- GOUVERNEMENT DU QUÉBEC (1991), *La protection sur mesure: un projet collectif*, Rapport du groupe de travail sur l'application des mesures de protection de la jeunesse, Québec, Ministère de la Santé et des Services sociaux.
- GOUVERNEMENT DU QUÉBEC (1998), *Pour une stratégie de soutien du développement des enfants et des jeunes: agissons en complices*, Québec, Ministère de la Santé et des Services sociaux.
- GREENE, J. C. (2006), « Toward a methodology of mixed methods social inquiry », *Research in the Schools*, 13(1), 93-98.
- GREENE, J. C., V. J. CARACELLI et W. F. GRAHAM (1989), « Toward a conceptual framework for mixed-method evaluation designs », *Educational Evaluation and Policy Analysis*, 11(3), 255-274.
- GREENHALGH, T. et al. (2004), *How to Spread Good Ideas: A Systematic Review of the Literature on Diffusion, Dissemination and Sustainability of Innovations in Health Service Delivery and Organization*, Rapport au NCCSDO.
- GREMBOWSKI, D. (2001), *The Practice of Health Program Evaluation*, Thousand Oaks, Sage Publications.

- GROLEAU, D., P. PLUYE et L. NADEAU (2007), « A mixed methods approach to the cultural understanding of distress and the non-use of mental health services », *Journal of Mental Health*, 16(6), 731-741.
- GROLEAU, R. (1990), « L'habileté à résoudre des problèmes interpersonnels chez des garçons qui présentent des troubles de comportement », mémoire de maîtrise, Université de Montréal.
- GUBA, E. G. et Y. A. LINCOLN (1987), « The countenances of fourth-generation evaluation: Description, judgment and negotiation », dans D. J. PALUMBO (dir.), *The Politics of Program Evaluation*, n° 15, Newbury Park, CA, Sage Publications, 202-234.
- GUBA, E. G. et Y. S. LINCOLN (1989), *Fourth Generation Evaluation*, Newbury Park, CA, Sage Publications.
- GUTSTEIN, S. E. (1987), « Family reconciliation as a response to adolescence crisis », *Family Process*, 26, 475-491.
- GUTSTEIN, S. E. et al. (1988), « Systemic crisis intervention as a response to adolescent crisis: An outcome study », *Family Process*, 27, 201-211.
- HACKING, I. (1999), *The Social Construction of What?*, Harvard University Press.
- HALL, P. (1993), « Policy paradigms, social learning and the state: The case of economic. Policymaking in Britain », *Comparative Politics*, 25, 275-296.
- HANSEN, M., N. NOHRIA et T. TIERNEY (1999), « What's your strategy for managing knowledge? », *Harvard Business Review*, mars-avril, 106-116.
- HAW, C. et L. ANDRES (1998), *ASK-ASSESS-ACT: Suicide Intervention Training Program for School Personnel*, Vancouver, British Columbia Council for Families.
- HEMPEL, C. G. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York, Free Press.
- HENRY, G. T. et M. M. MARK (2003), « Beyond use: Understanding evaluation's influence on attitudes and actions », *American Journal of Evaluation*, 24(3), 293-314.
- HITT, M. A. et al. (2001), « Direct and moderating effects of human capital on strategy and performance in professional service firms: A resource-based perspective », *Academy of Management Journal*, 44(1), 13-28.
- HOGWOOD B. W. et L. A. GUNN (1984), *Policy Analysis for the Real World*, Londres, Oxford University Press.
- HOUSE, E. R. (1980), *Evaluating with Validity*, Newbury Park, CA, Sage Publications.
- HOUSE, E. R. et K. R. HOWE (2000), « Deliberative democratic evaluation », *New Directions for Evaluation*, 85, printemps, 3-12.
- HUBERMAN, A. M. et M. B. MILES (1991), *Analyse des données qualitatives*, Bruxelles, De Boeck.

- HURTUBISE, R., M. VATZ LAAROUSSI et S. DUBUC (2000), *Jeunes de la rue et famille. Des productions sociales et des stratégies collectives au travers des mouvances du réseau*, Sherbrooke, Université de Sherbrooke.
- INPES (2002), *Évaluer l'éducation pour la santé: concepts et méthodes*, Paris, CFES, Comité régional d'Aquitaine d'éducation pour la santé, ENSP, Université Victor-Segalen Bordeaux 2.
- INSPQ (2001), *Le portrait de santé. Le Québec et ses régions*, Québec, Institut national de santé publique du Québec.
- INTERNATIONAL ORGANIZATION FOR COOPERATION IN EVALUATION (2006), *Creating and Developing Evaluation Organizations: Lessons Learned from Africa, Americas, Asia, Australasia and Europe*, Lima, Centro de Estudios y Promoción del Desarrollo.
- JACOB, S. (2005), « Réflexions autour d'une typologie des dispositifs institutionnels d'évaluation », *Canadian Journal of Program Evaluation*, 2092, 49-68.
- JACOB, S. (2006), « Trente ans d'évaluation de programmes au Canada: l'institutionnalisation interne en quête de qualité », *Revue française d'administration publique*, 119, 515-532.
- JACOB, S. (2007), « L'évaluation au cœur de la nouvelle gouvernance territoriale », dans S. Jacob, F. Varone et J.-L. Genard (dir.), *L'évaluation des politiques au niveau régional*, Bruxelles, Peter Lang, 179-197.
- JACOBS, S. et F. VARONE (2003), *Évaluer l'action publique: état des lieux et perspectives en Belgique*, Gent, Academia Press.
- JACOT, H. et al. (dir.) (2007), *L'écu, l'expert et le citoyen: vers un nouveau rôle des acteurs dans l'évaluation des politiques publiques*, Paris, L'Harmattan.
- JADAD, A. R. et al. (1996), « Assessing the quality of reports of randomized clinical trials: Is blinding necessary? », *Controlled Clinical Trials*, 17(1), 1.
- JAMES, P. D. (1989), *An Unsuitable Job for a Woman*, Londres, Penguin.
- JANOSZ, M. et al. (2007), *Rapport de validation du Questionnaire sur l'environnement socioéducatif des écoles primaires*, Groupe de recherche sur les environnements scolaires, Université de Montréal.
- JOHNSON, R. B., A. J. ONWUEGBUZIE et L. A. TURNER (2007), « Toward a definition of mixed methods research », *Journal of Mixed Methods Research*, 1(2), 112-133.
- JOINT COMMITTEE ON STANDARDS FOR EDUCATIONAL EVALUATION (JCSEE) (1981), *Standards for Evaluations of Educational Programs, Projects and Materials*, New York, McGraw-Hill.
- JONES, C. O. (1970), *An Introduction to the Study of Public Policy*, Belmont, Duxbury Press.
- KALLERUD, E. et I. RAMBERG (2002), « The order of discourse in surveys of public understanding of science », *Public Understanding of Science*, 11, 213-224.
- KATRAK, P. et al. (2004), « A systematic review of the content of critical appraisal tools », *BMC Medical Research Methodology*, 4(22).

- KAZI, M. A. F. (1996), « Single-case evaluation in the public sector using a combination of approaches », *Evaluation. The International Journal of Theory, Research and Practice*, 2(1), 85-97.
- KERLINGER, F. N. (1972), *Foundations of Behavioral Research*, New York, Holt, Rinhart & Winston.
- KHAN, K. S. et al. (2006), « WHO systematic review of causes of maternal deaths », *Lancet*, 367, 1066-1074.
- KINCHELOE J. L. et P. McLAREN (2000), « Rethinking critical theory and qualitative research », dans N. K. DENZIN et Y. S. LINCOLN (dir.), *Handbook of Qualitative Research*, 2^e éd., Thousand Oaks, Sage Publications, p. 282.
- KINGDON, J. W. (1984), *Agendas, Alternatives, and Public Policies*, Boston, Little Brown.
- KINNEY, J., D. HAAPALA et C. BOOTH (1991), *Keeping Families Together: The Homebuilders Model*, New York, Aldine de Gruyter.
- KINNEY, J., K. DITTMAR et W. FIRTH (1990), « Keeping families together: The homebuilders model », *Children Today*, 19, 14-19.
- KIRKPATRICK, D. L. (1996), *Evaluating Training Programs. The Four Levels*, San Francisco, Berrett-Koehler Publishers.
- KNOEFFEL, P., C. LARRUE et F. VARONE (2001), *Analyse et pilotage des politiques publiques*, Bâle, Helbing et Lichtenhahn.
- KOVNER, A. R. et T. G. RUNDALL (2006), « The promise of evidence-based management », *Frontiers of Health Services Management*, 22(3), 3-22.
- KRATHWOHL, D. R. (1982), *Social and Behavioral Science Research: A Framework for Conceptualizing, Implementing and Evaluating Research Studies*, San Francisco, Jossey-Bass.
- KUHN, T. (1983), *La structure des révolutions scientifiques*, Paris, Flammarion, trad. de *The Structure of Scientific Revolutions*, University of Chicago Press, 1962.
- LANE, P. J. et M. LUBATKIN (1998), « Relative absorptive capacity and interorganizational learning », *Strategic Management Journal*, 19(5), 461-477.
- LASSWELL, H. D. (1956), *The Decision Process: Seven Categories of Functional Analysis*, College Park, University of Maryland.
- LAST, J. M. (2004), *Dictionnaire d'épidémiologie*, Paris, Edisem/Maloine.
- LATOUR, B. (1989), *La science en action. Introduction à la sociologie des sciences*, Paris, Gallimard.
- LAVIS, J. (2004), « A political science perspective on evidence-based decision-making », dans L. LEMIEUX-CHARLES et F. CHAMPAGNE (dir.), *Using Knowledge and Evidence in Health Care. Multidisciplinary Perspectives*, Toronto, University of Toronto Press, 70-85.
- LECLERC, J. (2001), *Gérer autrement l'administration publique: la gestion par résultats*, 2^e éd., Sainte-Foy, Presses de l'Université du Québec.

- LECOMPTE, M. D. et J. PREISSE (1993), *Ethnography and Qualitative Design in Educational Research*, San Diego, Academic Press.
- LÉGARÉ, J. et A. DEMERS (dir.) (1993), *L'évaluation sociale : savoirs, éthique, méthodes*, Actes du 59^e congrès de l'ACSALF, Québec, Éditions du Méridien.
- LEMIEUX, V. (1995), *L'étude des politiques publiques : les acteurs et leur pouvoir*, Québec, Presses de l'Université Laval.
- LEMIEUX, V. (2002), *L'étude des politiques publiques : les acteurs et leur pouvoir*, 2^e éd., Sainte-Foy, Presses de l'Université Laval.
- LEMIEUX, V. (2006), « Évaluation de programmes et analyse des politiques », *Télescope : Revue d'analyse comparée en administration publique*, 13(1), 1-8.
- LEMIEUX-CHARLES, L. et F. CHAMPAGNE (dir.) (2004), *Using Knowledge and Evidence in Health Care: Multidisciplinary Perspectives*, Toronto, University of Toronto Press.
- LEVIN-ROZALIS, M. (1987), « The process of partially evaluating the change caused by intervention in a community », thèse de maîtrise en sciences sociales, Université Ben-Gourion du Néguev (en hébreu).
- LEVIN-ROZALIS, M. (1998), « Are evaluation and research really the same thing? », *Megamot*, 39(3), 303-319 (en hébreu).
- LEVIN-ROZALIS, M. (2000), « Abduction, a logical criterion for program evaluation », *Evaluation. The International Journal of Theory, Research and Practice*, 6(4), 411-428.
- LEVIN-ROZALIS, M. et N. BAR-ON (1994), *Diagnosis and Evaluation in Youth Village « Eshel-Hanasi »*, Internal report submitted to the youth village board, management and staff (en hébreu).
- LEVITON, L. C. (2003), « Evaluation use : Advances, challenges and applications », *American Journal of Evaluation*, 24(4), 525-535.
- LI, X. F. et al. (1996), « The post-partum period : The key to maternal mortality », *International Journal of Gynæcology and Obstetrics*, 54, 1-10.
- LINCOLN, Y. S. et E. G. GUBA (1986), « But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation », *New Directions for Program Evaluation*, juin, 73-81.
- LINDBLOM, C. E. (1979), « Still muddling, not yet through », *Public Administration Review*, 39(6), 517-525.
- LINDBLOM, C. E. et D. K. COHEN (1979), *Usable Knowledge : Social Science and Social Problem Solving*, New Haven et Londres, Yale University Press.
- LOMAS, J. et al. (2005), *Conceptualizing and Combining Evidence for Health System Guidance*, Rapport de la Canadian Health Services Research Foundation.
- MAINE, D. et al. (1997), *La conception et l'évaluation des programmes de prévention de la mortalité maternelle*, New York, Colombia University.

- MANNING-KENDRICK, J. (1991), «Crisis intervention in child abuse: A family treatment approach», dans A. R. ROBERTS (dir.), *Contemporary Perspectives on Crisis Intervention and Prevention*, Englewood Cliffs, Prentice Hall.
- MARK, M. M., G. T. HENRY et G. JULNES (2000), *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*, San Francisco, Jossey-Bass.
- MATHISON, S. (dir.) (2004), *Encyclopedia of Evaluation*, Thousand Oaks, Sage Publications.
- MAYS, N. et C. POPE (2000), «Qualitative research in health care: Assessing quality in qualitative research», *British Medical Journal*, 320(7226), 50-52.
- MAYS, N., C. POPE et J. POPAY (2005), «Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field», *Journal of Health Services Research and Policy*, 190(3), Suppl. 1, 6-20.
- MCDONALD, B., P. ROGERS et B. KEFFORD (2003), «Teaching people to fish? Building the evaluation capability of public sector organizations», *Evaluation*, 9(1), 9-29.
- MILES, M. et M. HUBERMAN (1994), *Qualitative Data Analysis*, Newbury Park, CA, Sage Publications.
- MILSTEIN, B. et D. A. COTTON (2000), «Defining concepts for the presidential strand on building evaluation capacity», *American Evaluation Association*. En ligne : www.eval.org/eval2000/public/presstrand.pdf
- MINISTÈRE DE L'ÉCONOMIE ET DES FINANCES DU MALI (2002), *Politique nationale de Population-Révision 1*, Bamako, Ministère de l'Économie et des Finances.
- MINTZBERG, H. (1979), *The Structuring of Organizations: A Synthesis of the Research*, Englewood Cliffs, Prentice-Hall.
- MOHER, D. et al. (1999), «Assessing the quality of reports of randomised trials: Implications for the conduct of meta-analyses», *Health Technology Assessment*, 3(12), 1-108.
- MONNIER, E. (1992), *Évaluations de l'action des pouvoirs publics: du projet au bilan*, Paris, Économica.
- MOORE, D., D. GAGNON et A. PERREAU (1998), *Stratégie de travail de rue. Pertinence et recommandations*, Longueuil, Régie régionale de la santé et des services sociaux Montérégie, 37.
- MOR, Y. (1992), «The use of grey documentation in the comprehensive evaluation of a broadly aimed social program», Paper presented at the International Expert Meeting on Information and Social Welfare Policy and Planning, novembre, Jérusalem.
- MORELL, J. (2010), *Evaluating in the Face of Uncertainty*, New York, Guilford Press.
- MORGAN, D. L. (2007), «Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods», *Journal of Mixed Methods Research*, 1(1), 48-76.

- MUCHIELLI, A. (1994), *Les méthodes qualitatives*, Paris, Presses universitaires de France.
- MULLER, P. (2003), *Les politiques publiques*, Paris, Presses universitaires de France.
- NACHEMIAS, D. et H. NACHEMIAS (1982), *Research Methods in Social Sciences*, Tel Aviv, Am-Oved (en hébreu).
- NAHAPIET, J. et S. GHOSHAL, (1998), « Social capital, intellectual capital, and the organizational advantage », *Academy of Management Review*, 23(2), 242-266.
- NELSON, K. E., M. J. LANDSMAN et W. DEUTELBAUM (1990), « Three models of family-centered placement prevention services », *Child Welfare*, 114, 1-21.
- NEVIS, C. E., A. J. DIBELLA et J. M. GOULD (1995), « Understanding organizations as learning systems », *Sloan Management Review*, 36(2), 73.
- NEVO, D. (1989), *Useful Evaluation: Evaluating Educational and Social Projects*, Tel Aviv, Massada (en hébreu).
- NINACS, W. A. (2002), « Types et processus d'empowerment dans les initiatives de développement économique communautaire au Québec », thèse de doctorat, Québec, Université Laval, École de service social.
- NONAKA, I. (1994), « A dynamic theory of organizational knowledge creation », *Organization Science*, 5(1), 14-37.
- O'CATHAIN, A. (2010), « Assessing the quality of mixed methods research: Towards a comprehensive framework », dans A. TASHAKKORI et C. TEDDLIE (dir.), *Handbook of Mixed Methods in Social and Behavioral Research*, 2^e éd., Thousand Oaks, Sage Publications, 531-555.
- ORGANISATION MONDIALE DE LA SANTÉ (2008), *Mortalité maternelle en 2005: Estimations OMS, UNICEF, FNUAP et Banque mondiale*, Genève, OMS.
- OWEN, M. J. et P. J. ROGERS (1999), *Program Evaluation: Forms and Approaches*, Londres, Sage Publications.
- PAGE, R. et al. (2012), « Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review », *International Journal of Nursing Studies*, 49, 47-53.
- PAQUIN, S., L. BERNIER et G. LACHAPPELLE (dir.) (2011), *L'analyse des politiques publiques*, Montréal, Presses de l'Université de Montréal.
- PARAZELLI, M. (2002), *La rue attractive. Parcours et pratiques identitaires des jeunes de la rue*, Sainte-Foy, Presses de l'Université du Québec.
- PATTON, M. Q. (1978), *Utilization-Focused Evaluation*, Newbury Park, CA, Sage Publications.
- PATTON, M. Q. (1981), *Creative Evaluation*, Newbury Park, CA, Sage Publications.
- PATTON, M. Q. (1988a), « The evaluator's responsibility for utilization », *Evaluation Practice*, 9(2), 5-24.
- PATTON, M. Q. (1988b), « How primary is your identity as an evaluator? », *Evaluation Practice*, 9(2), 87-92.

- PATTON, M. Q. (1994), « Developmental evaluation », *Evaluation Practice*, 15(3), 311-320.
- PATTON, M. Q. (1996), « A world larger than formative and summative », *Evaluation Practice*, 17(2), 131-144.
- PATTON, M. Q. (1997), *Utilization-Focused Evaluation: The New Century Text*, 3^e éd., Thousand Oaks, Sage Publications.
- PATTON, M. Q. (1997b), « Implementation evaluation: What happened in the program? », dans *Utilization-Focused Evaluation: The New Century Text*, 3^e éd., Thousand Oaks Sage Publications, 195-214.
- PATTON, M. Q. (1998), « Discovering process use », *Evaluation*, 4(2), 225-233.
- PATTON, M. Q. (2000), « Utilization-focused evaluation », dans D. L. STUFFLEBEAM, G. F. MADAUS et T. KELLAGHAN (dir.), *Evaluation Models*, Boston, Kluwer.
- PATTON, M. Q. (2001), « Evaluation, knowledge management, best practices, and high quality lessons learned », *American Journal of Evaluation*, 22(3), 329-336.
- PATTON, M. Q. (2002), *Qualitative Research and Evaluation Methods*, Thousand Oaks, Sage Publications.
- PAWSON, R. (1996), « Three steps to constructivist heaven », *Evaluation*, 2(2), 213-219.
- PAWSON, R. (2006), *Evidence-Based Policy: A Realist Perspective*, Londres, Sage Publications.
- PAWSON, R. et al. (2005), « Realist review: A new method of systematic review designed for complex policy interventions », *Journal of Health Services Research and Policy*, 10(3), Suppl. 1, 21-34.
- PAWSON, R. et N. TILLEY (1997), *Realistic Evaluation*, Londres, Sage Publications.
- (2005), « Realistic evaluation », dans S. MATHISON (dir.), *Encyclopedia of Evaluation*, Thousand Oaks, Sage Publications, 362-637.
- PAXTON, A. et al. (2005), « The evidence for emergency obstetric care », *International Journal of Gynecology & Obstetrics*, 88, 181-193.
- PEIRCE, C. S. (1960), *Collected Papers of Charles Sanders Peirce*, vol. 3-4, Cambridge, Belknap Press.
- PÉLADÉAU, N. et C. Mercier (1993), « Approches qualitative et quantitative en évaluation de programmes », *Sociologie et sociétés*, 25(2), 111-124.
- PENROSE, E. T. (1959), *The Theory of the Growth of the Firm*, Oxford, Blackwell.
- PERES Y. et G. YATSIV (1994), *Introduction to Research Methods in Behavioral Sciences*, 2^e éd., Jérusalem, Academon Press (en hébreu).
- PERRET, B. (2001), *L'évaluation des politiques publiques*, Paris, La Découverte.
- PFEFFER, J. et R. I. SUTTON (1999), *The Knowing-Doing Gap: How Smart Companies Turn Knowledge into Action*, Cambridge, Harvard Business School Press.

- PIERSON, P. (2000), « Increasing returns, path dependence, and the study of politics », *American Political Science Review*, 94(2), 251-267.
- PINEAULT, R. et C. DAVELUY (1995), *La planification de la santé : concepts, méthodes, stratégies*, Montréal, Agence d'Arc.
- PITTMAN, F. S. (1987), *Turning Points: Treating Families in Transition and Crisis*, New York, W. W. Norton.
- PLANTE, J. (1994), *Évaluation de programmes*, Sainte-Foy, Presses de l'Université Laval.
- PLUYE, P. et al. (2005), « Program sustainability begins with the first events », *Evaluation and Program Planning*, 28(2), 123-137.
- PLUYE, P. et al. (2007), « Mixed studies reviews in health sciences », article présenté à la Mixed Methods Conference, Cambridge.
- PLUYE, P. et al. (2009), « Understanding divergence of quantitative and qualitative data (or results) in mixed methods studies », *International Journal of Multiple Research Approaches*, 3(1).
- PLUYE, P. et al. (2011), Proposal: A mixed methods appraisal tool for systematic mixed studies reviews. Retrieved from <http://mixedmethodsappraisaltool-public.pbworks.com>. Archived by WebCite® at www.webcitation.org/5tTRTc9yJ
- PLUYE, P., L. POTVIN et J.-L. DENIS (2004), « Making public health programs last. Conceptualizing sustainability », *Evaluation and Program Planning*, 27(2), 121-133.
- POLAND, B. D., E. TUPKER et K. BRELAND (2002), « Involving street youth in peer harm reduction education : The challenges of evaluation », *Canadian Journal of Public Health*, 93(5), 344-348.
- POLANYI, M. (1966), *The Tacit Dimension*, Londres, Routledge and Kegan Paul.
- POPAY, J., A. ROGERS et G. WILLIAMS (1998), « Rationale and standards for the systematic review of qualitative literature in health services research », *Qualitative Health Research*, 8(3), 341-351.
- POPPER, K. R. (1959), *The Logic of Scientific Discovery*, Londres, Hutchinson and Co.
- POSAVAC, E. J. et R. G. CAREY (1992), *Program Evaluation: Methods and Case Studies*, 2^e éd., Englewood Cliffs, Prentice-Hall.
- PRESKILL, H., B. ZUCKERMAN et B. MATTHEWS (2003), « An exploratory study of process use : Findings and implications for future research », *American Journal of Evaluation*, 24(4), 423-442.
- PRESSMAN, J. L. et A. WILDAVSKY (1973), *Implementation*, Berkeley, University of California Press.
- PROJET OBJECTIFS DU MILLÉNAIRE (2005), *Investir dans le développement : plan pratique pour réaliser les objectifs du Millénaire pour le développement*, ONU.
- QUINN J. B. (1992), *Intelligent Enterprise*, New York, Free Press.

- QUINN, J. B., P. ANDERSON et S. FILKEILSTEIN (1996), « Managing professional intellect: Making the most of the best », *Harvard Business Review*, 74(2), 71.
- RALLIS, S. et G. ROSSMAN (2003), « Mixed methods in evaluation contexts: A pragmatic framework », dans A. TASHAKKORI et C. TEDDLIE (dir.), *Handbook of Mixed Methods in Social and Behavioral Research*, Thousand Oaks, Sage Publications, 491-512.
- RENZULLI, J. S. (1975), *A Guidebook for Evaluation Programs for the Gifted and Talented*, Ventura, Office of the Ventura County Superintendent of Schools.
- RIDDE, V. (2005), « S'informer avant de réaliser pour la première fois une évaluation de programmes en promotion de la santé », *Reviews of Health Promotion and Education Online*, www.rhpeo.org/reviews/2005/30/index.htm
- RIDDE, V. (2006a), « Programmes communautaires et innovations méthodologiques : participation, accompagnement et empowerment », *Revue canadienne d'évaluation de programmes*, 21(3), 133-136.
- RIDDE, V. (2006b), « Suggestions d'amélioration d'un cadre conceptuel de l'évaluation participative », *Revue canadienne d'évaluation de programmes*, 21(2), 1-23.
- RIDDE, V. et al. (2003), « L'évaluation participative de type empowerment : une stratégie pour le travail de rue » *Service social*, 50(1), 263-279.
- RIDDE, V. et al. (2006), « Le magicien du temps : approche participative axée sur le développement d'un projet et l'utilisation des résultats d'une évaluation », *Revue canadienne d'évaluation de programmes*, 21(3), 235-255.
- RIDDE, V. et al. (2009), Programme evaluation training for health professionals in francophone Africa: Process, competence acquisition and use », *Human Resource for Health*, 15 janvier, 7(1), 3. doi:10.1186/1478-4491-7-3.
- RIDDE, V. et S. SAHIBULLAH (2005), « Evaluation capacity building and humanitarian organization », *Journal of MultiDisciplinary Evaluation*, 3, 78-112.
- RIDDE, V., J. POMMIER et F. JABOT (2010), « De l'évaluation des programmes à la réduction des inégalités sociales de santé dans L. Potvin (dir.), *Inégalités sociales de santé. Guide d'aide à l'action*, Paris, INPES.
- RIDDE, V., J. POMMIER et F. JABOT (2009), « De l'évaluation des programmes à la réduction des inégalités sociales de santé », dans L. POTVIN et al. (dir.), *Inégalités sociales de santé : guide d'aide à l'action*, Montréal, INPES, coll. « Varia ».
- RIDDE, V., P. PLUYE et L. QUEUILLE (2006), « Évaluer la pérennité des programmes de santé publique. Un outil et son application en Haïti », *Revue d'épidémiologie et de santé publique*, 54(5), 421-431.
- RIDDE, V., T. DELORMIER et L. GAUDREAU (2007), « Evaluation of empowerment and effectiveness: Universal concepts? », dans D. V. McQUEEN et C. M. JONES (dir.), *Global Perspectives on Health Promotion Effectiveness*, New York, Springer Science and Business Media, p. 389-403.

- ROBERTS, A. R. (1990), « An overview of crisis theory and crisis intervention », dans A. R. ROBERTS (dir.), *Crisis Intervention Handbook, Assessment, Treatment and Research*, Belmont, Wadsworth.
- ROBERTS, A. R. (1991), « Conceptualizing crisis theory and the crisis intervention model », dans A. R. ROBERTS (dir.), *Contemporary Perspectives on Crisis Intervention and Prevention*, Englewood Cliffs, Prentice Hall.
- RONSMANS, C. et W. J. GRAHAM (2006), « Maternal mortality: Who, when, where and why », *Lancet*, 368, 1189-1200.
- ROOTMAN, I. et al. (2001), « A framework for health promotion evaluation », dans I. ROOTMAN et al. (dir.), *Evaluation in Health Promotion: Principles and Perspectives*, Copenhagen, WHO Regional Office for Europe, 7-39.
- ROSE, R. et P. L. DAVIES (1993), *Inheritance in Public Policy: Change without Choice in Britain*, Yale University Press.
- ROSSI, P. H. et H. E. FREEMAN (1982), *Evaluation: Systemic Approach*, Newbury, CA, Sage Publications.
- ROSSI, P. H. et J. D. RIGHT (1984), « Evaluation research: An assessment », *Annual Review of Sociology*, 10, 331-352.
- ROSSI, P. H., M. W. LIPSEY et H. E. FREEMAN (2004), « Assessing and monitoring program process », dans *Evaluation: A Systematic Approach*, 7^e éd., Thousand Oaks, Sage Publications, 169-199.
- ROUGE, J.-C. (2004), « The origin and development of the African evaluation guidelines », *New Directions for Evaluation*, 104, 55-66.
- ROUSSEAU, D. M. et S. MCCARTHY (2007), « Educating managers from an evidence-based perspective », *Academy of Management Learning and Education*, 6(1), 84-101.
- RYCUS, J. S. et R. C. HUGHES (2000), *What Is Competency-based Inservice Training?*, Columbus, Institute for Human Services.
- SABAR, N. (1990), *Qualitative Research in Teaching and Learning*, Tel Aviv, Mas-sada Press (en hébreu).
- SABATIER, P. A. (1997), « The status and development of policy theory », *Policy Currents*, 7(4), 1-10.
- SAINT-MARTIN, D. (2004), « Managerialist advocate or “control freak”? The Janus-faced Office of the Auditor General », *Canadian Public Administration*, 47(2), 121-140.
- SALTMAN, R., R. BUSSE et E. MOSSIALIS (dir.) (2002), *Regulating Entrepreneurial Behaviour in European Health Care Systems*, Buckingham, Open University Press.
- SANDELOWSKI, M. (1993), « Rigor or rigor mortis: The problem of rigor in qualitative research revisited », *Advances in Nursing Science*, 16(2), 1-8.

- SANDELOWSKI, M., C. I. VOILS et J. BARROSO (2007), «Comparability work and the management of difference in research synthesis studies», *Social Science and Medicine*, 64(1), 236-247.
- SANDERSON, S., I. D. TATT et J. P. T. HIGGINS (2007), «Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography», *International Journal of Epidemiology*, 36(3), 666-676.
- SCHNEIDER, A. L. (1986), «Evaluation research and political science: An argument against the division of scholarly labor», *Policy Studies Review*, 6(2), 222-231.
- SCHORR, L. B. et D. SCHORR (1988), *Within Our Reach: Breaking the Cycle of Disadvantage*, New York, Anchor Books.
- SCHWANDT, T. A. (2002), «Democratization, globalization, and performativity: Implications for learning in and from evaluation», Discours d'ouverture prononcé au 5^e Congrès biennal de l'European Evaluation Society, 12 octobre 2002, Séville, Espagne.
- SCRIVEN, M. (1967), «The methodology of evaluation», dans R. E. STAKE (dir.), *Curriculum Evaluation*, AERA Monograph Series on Evaluation, Chicago, Rand McNally.
- SCRIVEN, M. (1983), «The evaluation taboo», *New Direction in Program Evaluation*, 75-82.
- SCRIVEN, M. (1991), «Beyond formative and summative evaluation», dans M. W. McLAUGHLIN et D. C. PHILLIPS (dir.), *Evaluation and Education: At Quarter Century*, 90th Yearbook of the National Society for the Study of Education, University of Chicago Press, 18-64.
- SCRIVEN, M. (1991), «Beyond formative and summative evaluation», dans M. W. McLAUGHLIN et D. C. PHILLIPS (dir.), *Evaluation and Education: At Quarter Century*, University of Chicago Press, 18-64.
- SCRIVEN, M. (1991a), *Evaluation Thesaurus*, 4^e éd., Newbury, CA, Sage Publications.
- SCRIVEN, M. (1997), «Empowerment evaluation examined», *Evaluation Practice*, 18(2), 165-175.
- SEGSWORTH, R. V. (2002), «Evaluation in the twenty-first century: Two perspectives on the canadian experience», dans J.-E. FURUBO *et al.* (dir.), *International Atlas of Evaluation*, New Brunswick, Transaction Publishers, 175-189.
- SEGSWORTH, R. V. (2005), «Program evaluation in the gouvernement of Canada: Plus ça change...», *The Canadian Journal of Program Evaluation*, 29(3), 175-197.
- SHADISH, W. R., T. D. COOK et D. T. CAMPBELL (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Boston, Houghton Mifflin.

- SHAFFER, D. et al. (2001), « Practice parameters for the assessment and treatment of children and adolescents with suicidal behaviour », *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(7), 24-51.
- SHAFFER, D. et L. CRAFT (1999), « Methods of adolescent suicide prevention », *Journal of Clinical Psychiatry*, 60, Suppl. 2, 70-74.
- SHANK, G. et D. J. CUNNINGHAM (1996), *Modelling the Six Modes of Peircean Abduction for Educational Purposes*, www.cs.indiana.edu/event/maics96/Proceedings/shank.html
- SHARP, L. et J. FRECHTLING (1997), « Introducing this handbook: The need for a handbook on designing and conducting mixed method evaluations », dans *User-Friendly Handbook for Mixed Method Evaluations*, Arlington, National Science Foundation, Directorate for Education and Human Resources.
- SHAW, I., J. C. GREENE et M. M. MARK (dir.) (2006), *The Sage Handbook of Evaluation*, Thousand Oaks, Sage Publications.
- SHULHA, L. M. et J. B. COUSINS (1997), « Evaluation use: Theory, research, and practice since 1986 », *Evaluation Practice*, 18(3), 195-208.
- SIMARD, P. et al. (2002), *Le travail de rue à Rouyn-Noranda : une histoire de collaboration*, Rouyn-Noranda, Arrimage Jeunesse, Régie régionale de la santé et des services sociaux de l'Abitibi-Témiscamingue.
- SIMARD, P., J. CARON et J.-P. MARQUIS (2000), *Évaluation du programme « travail de rue » de Rouyn-Noranda : cartographie dynamique du jeu des acteurs et des actrices*, Rouyn-Noranda, Régie régionale des services sociaux de l'Abitibi-Témiscamingue.
- SIMON, H. A. (1945), *Administrative Behavior*, New York, Macmillan.
- SLIVE, A. B. et N. MCCONKEY-RADETZKI (1989), « The utilization of residential treatment as family crisis intervention », *Journal of Child Care*, 3, 59-68.
- SOCIÉTÉ CANADIENNE D'ÉVALUATION (1999), *Série des compétences essentielles*, Ottawa, Société canadienne d'évaluation, www.evaluationcanada.ca
- STAKE, R. E. (1969), « Language, rationality and assessment », dans H. W. BEATY (dir.), *Improving Educational Assessment and an Inventory of Measures of Affective Behavior*, Washington, Association of Supervision and Curriculum Development.
- STAKE, R. E. (2004), *Standards-Based and Responsive Evaluation*, Thousand Oaks, Sage Publications.
- STEELE, W. et M. RAIDER (1991), *Working with Families in Crisis: School-Based Intervention*, New York, Guilford.
- STEVAHN, L. et al. (2005), « Establishing essential competencies for program evaluators », *American Journal of Evaluation*, 26(1), 43-59.
- STONE, D. A. (1989), « Causal stories and the formation of policy agendas », *Political Science Quarterly*, 104(2), 281-300.

- STRAUSS, A. (1987), *Qualitative Analysis for Social Scientists*, New York, Cambridge University Press.
- STUART, C., J. K. WAALEN et E. HAELSTROMM (2003), « Many helping hearts: An evaluation of peer gatekeeper training in suicide risk assessment », *Death Studies*, 27, 321-333.
- STUFFLEBEAM, D. L. (1971), « The CIPP model for program evaluation », dans G. E. MADAUS, M. SCRIVEN et D. L. STUFFLEBEAM (dir.), *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, Boston, Kluwer-Nijhoff.
- STUFFLEBEAM, D. L. (2001), « Evaluation models », *New Directions for Evaluation*, n° 89, San Francisco, Jossey-Bass.
- STUFFLEBEAM, D. L. et al. (1974), *Educational Evaluation and Decision Making*, 4^e éd., Itaska, F. E. Peacock Publishers.
- SUGARMAN, S. et C. MASHESTER (1985), « The family crisis intervention literature: What is meant by "family"? », *Journal of Marital and Family Therapy*, 2, 167-177.
- TASHAKKORI, A. et C. TEDDLIE (2003), *Handbook of Mixed Methods in Social and Behavioral Research*, Thousand Oaks, Sage Publications.
- TEDDLIE, C. et A. TASHAKKORI (2003), « Major issues and controversies in the use of mixed methods in the social and behavioural sciences », dans A. TASHAKKORI et C. TEDDLIE (dir.), *Handbook of Mixed Methods in Social and Behavioral Research*, Thousand Oaks, Sage Publications, 3-50.
- TEECE, D. J. (1981), « The market for know-how and the efficient international transfer of technology », *Annals of the American Academy of Political and Social Science*, 458, 81-96.
- THADDEUS, S. et D. MAINE (1994), « Too far to walk: Maternal mortality in context », *Social Science and Medicine*, 38(8), 1091-1110.
- THONNEAU, P. et al. (2002), « Abortion and maternal mortality in Africa », *New England Journal of Medicine*, 347, 24.
- TIERNEY, R. J. (1988), « Comprehensive evaluation for suicide intervention training », thèse de doctorat, Calgary, University of Calgary.
- TIERNEY, R. J. (1994), « Suicide intervention training evaluation: A preliminary report », *Crisis*, 15(2), 69-76.
- TOUSIGNANT, R. et D. MORISSETTE (1990), *Les principes de la mesure et de l'évaluation des apprentissages*, 2^e éd., Boucherville, Gaëtan Morin éditeur.
- TREMBLAY, R. E. et al. (1992), « A prosocial scale for the preschool behaviour questionnaire: Concurrent and predictive correlates », *International Journal of Behavioral Development*, 15(2), 227-245.
- TROCHIM, W. M. (1985), « Pattern matching, validity, and conceptualization in program evaluation », *Evaluation Review*, 9, 575-604.

- TROSA, S. (2003), *Le guide de la gestion par programme: vers une culture du résultat*, Paris, Éditions d'Organisation.
- TUCKER, A. S. et J. V. DEMPSEY (1991), « A semiotic model for program evaluation », *The American Journal of Semiotics*, 8(4), 73-103.
- TURNER, J. H. (1986), *The Structure of Sociological Theory*, 4^e éd., New York, Wadsworth.
- TYLER, R. W. (1942), « General statements on evaluation », *Journal of Educational Research*, 35(7), 492-501.
- TYLER, R. W. (1950), *Basic Principles of Curriculum and Instruction*, Chicago, University of Chicago Press.
- UMANA, R., S. J. GROSS et M. T. MCCONVILLE (1980), *Crisis in the Family: Three Approaches*, New York, Gardner Press.
- WALLACE, L. W. (1971), *Sociological Theory*, Londres, Heinemann Educational Books.
- WALSHE, K. et T. G. RUNDALL (2001), « Evidence-based management: From theory to practice in health care », *Milbank Quarterly*, 79(3), 429.
- WEAVER, L. et J. B. COUSINS (2005), « Unpacking the participatory process », *Journal of Multidisciplinary Evaluation*, 1, 19-40.
- WEBER, M. (1995), *Économie et société: les catégories de la sociologie*, Paris, Plon.
- WEISS, C. H. (1988a), « Evaluation for decisions: Is anybody there? Does anybody care? », *Evaluation Practice*, 9(1), 5-19.
- WEISS, C. H. (1988b), « If program decisions hinged only on information: A response to Patton », *Evaluation Practice*, 9(3), 5-19.
- WEISS, C. H. (1996), « Theory based evaluation: Past, present and future », Communication présentée au congrès annuel de l'American Evaluation Association, Atlanta, Californie.
- WEISS, C. H. (1998), « Have we learned anything new about the use of evaluation? », *American Journal of Evaluation*, 19(1), 21-33.
- WEISS, C. H. (1998), *Evaluation*, Up Saddle River, NJ, Prentice Hall.
- WHITMORE, E. et C. MCKEE (2001), « Six street youth who could... », dans P. REASON et H. BRADBURY (dir.), *Handbook of Action Research. Participative Inquiry and Practice*, Thousand Oaks, CA, Sage Publications, 396-402.
- WIENER, R. L. et al. (1994), « Needs assessment. Combining qualitative interviews and concept mapping methodology », *Evaluation Review*, 18, 227-240.
- WILDAVSKY, A. (1979), *Speaking Truth to Power: The Art and Craft of Policy Analysis*, Boston, Little and Brown.
- WILLIAM, P. et al. (2001), « Knowing the risks: Theory and practice in financial market trading », *Human Relations*, 54(7), 887-910.

- WILLIAMS, B. (2001), « Building evaluation capability ». En ligne: <http://users.actrix.co.nz/bobwill/>
- WILSON, H. S. et S. A. HUTCHISON. (1991), « Triangulation of qualitative methods: Heideggerian hermeneutics and grounded theory », *Qualitative Health Research*, 1, 263.
- YIN, R. K. (1994), *Case Study Research: Design and Methods*, Londres, New Delhi, Sage Publications.
- ZAHRA, S. A. et G. GEORGE (2002), « Absorptive capacity: A review, reconceptualization, and extension », *Academic Management Review*, 27(2), 185-203.
- ZORZI, R., M. MCGUIRE et B. PERRIN (2002), *Projet de la société canadienne d'évaluation visant à appuyer la défense des intérêts et le perfectionnement professionnel*, Toronto, Zorzi & Associates.

Les auteurs

ARISTIDE BADO, Institut de recherche en sciences de la santé, Ouagadougou, Burkina Faso

MICHAEL BAMBERGER, Independent consultant at Development Evaluation, Portland, Oregon

MURIELLE BAUCHET, Consultante en santé publique, France

DIANE BERTHELETTE, Département d'organisation et ressources humaines - École des sciences de la gestion, Université du Québec à Montréal

PIERRE BLAISE, Direction régionale des affaires sanitaires et sociales, Pays de la Loire, Nantes, France

FRANÇOIS BOWEN, Faculté des sciences de l'éducation, Université de Montréal

FRANÇOIS CHAGNON, Département de psychologie, Université du Québec à Montréal

NADIA CUNDEN, Institute of Applied Health Sciences, Foresterhill, University of Aberdeen

CHRISTIAN DAGENAIS, Département de psychologie, Université de Montréal

PIERRE-MARC DAIGNEAULT, Département de science politique, Université Laval

LUC DESNOYER, Département des sciences biologiques, Université du Québec à Montréal

DIDIER DUPONT, Département de psychologie, Université de Montréal

JULIE DUTIL, Centre de liaison sur l'intervention et la prévention psychosociales (CLIPP)

FRANÇOISE FORTIN, École de psychoéducation, Université de Montréal

PIERRE FOURNIER, Centre de recherche du Centre hospitalier de l'Université de Montréal (CRCHUM) et Département de médecine sociale et préventive, Université de Montréal

MARIE GERVAIS, Département de médecine sociale et préventive, Université Laval

ANNE GUICHARD, Institut national de prévention et d'éducation à la santé, France

ROBERT R. HACCOUN, Département de psychologie, Université de Montréal

JANIE HOULE, Département de psychologie, Université du Québec à Montréal

FRANÇOISE JABOT, École des hautes études en santé publique, France

STEVE JACOB, Département de science politique, Université Laval

KADIDIATOU KADIO, Institut de recherche en sciences de la santé, Ouagadougou, Burkina Faso

SENI KOUANDA, Institut de recherche en sciences de la santé, Ouagadougou, Burkina Faso

FRANCINE LABOSSIERE, Agence de la santé publique du Canada

MIRI LEVIN-ROZALIS, Department of Education, Ben-Gurion University, Israël

ISABELLE MARCOUX, École interdisciplinaire des sciences de la santé, Université d'Ottawa

PIERRE McDUFF, Département de psychologie, Université de Montréal

FRÉDÉRIC NAULT-BRIÈRE, Département de psychologie, Université de Montréal

BERNARD PERRET, Conseil général de l'environnement et du développement durable, France

PIERRE PLUYE, Département de médecine familiale, Université McGill

NANCY L. PORTEOUS, Health Canada and Public Health Agency of Canada

MICHAEL QUINN PATTON, independent organizational development and program evaluation consultant, St. Paul, Minnesota

VALÉRY RIDDE, Centre de recherche du Centre hospitalier de l'Université de Montréal (CRCHUM) et Département de médecine sociale et préventive, Université de Montréal

ÉMILIE ROBERT, Centre de recherche du Centre hospitalier de l'Université de Montréal (CRCHUM) et Département de médecine sociale et préventive, Université de Montréal

PATRICIA ROGERS, Global, Urban and Social Studies, RMIT University, Melbourne, Australia

CHRISTINE ROTHMAYR, Département de science politique, Université de Montréal

JIM RUGH, Independent international evaluation consultant, Sevierville, Tennessee

CAROLINE TOURIGNY, Centre de recherche du Centre hospitalier de l'Université de Montréal (CRCHUM)

JOSEFIEN VAN OLMEN, Département de santé publique, Institut de médecine tropicale, Anvers, Belgique

SOPHIE WITTER, Institute of Applied Health Sciences, Foresterhill, University of Aberdeen

MAURICE YAMEOGO, Institut de recherche en sciences de la santé, Ouagadougou, Burkina Faso

ROBERT K. YIN, American University's School of International Service, Washington D.C.

Table des matières

Présentation	7

PREMIÈRE PARTIE CONCEPTS ET APPROCHES

1	Introduction générale à l'évaluation de programmes <i>Valéry Ridde et Christian Dagenais</i>	13
2	Recherche et évaluation de programmes <i>Miri Levin-Rozalis</i>	33
3	La construction d'un jugement <i>Bernard Perret</i>	53
4	L'analyse des politiques publiques <i>Steve Jacob et Christine Rothmayr</i>	71
5	La construction du modèle logique d'un programme <i>Nancy L. Porteous</i>	89
6	Attribution et causalité des effets <i>Robert R. Haccoun et Pierre McDuff</i>	109
7	Les méthodes mixtes <i>Pierre Pluye</i>	125
8	L'évaluation axée sur l'utilisation <i>Michael Quinn Patton et Francine LaBossière</i>	145
9	Une stratégie pour composer avec les contraintes inhérentes à la pratique <i>Michael Bamberger et Jim Rugh</i>	161

10	Théorie et pratiques des études de cas en évaluation de programmes	179
	<i>Robert K. Yin et Valéry Ridde</i>	
11	Une introduction à l'évaluation économique des programmes	199
	<i>Sophie Witter et Nadia Cunden</i>	
12	Le renforcement des capacités en évaluation	213
	<i>Patricia Rogers et Marie Gervais</i>	
13	Conceptualiser et mesurer la participation à l'évaluation	233
	<i>Pierre-Marc Daigneault et Steve Jacob</i>	
14	Théorie et pratique de l'approche <i>Realist</i> pour l'évaluation des programmes	255
	<i>Valéry Ridde, Émilie Robert, Anne Guichard, Pierre Blaise et Josefien Van Olmen</i>	

DEUXIÈME PARTIE

PRATIQUES ÉVALUATIVES

15	Une évaluation participative avec renforcement du pouvoir d'agir	279
	<i>Valéry Ridde</i>	
16	Une évaluation des besoins en formation d'intervenants en prévention	295
	<i>François Chagnon, Janie Houle et Isabelle Marcoux</i>	
17	Une évaluation des processus d'un programme de protection de la jeunesse	309
	<i>Christian Dagenais et Didier Dupont</i>	
18	Une évaluation des processus d'un programme de référence-évacuation au Mali	325
	<i>Pierre Fournier, Caroline Tourigny et Francine LaBossière</i>	
19	Une évaluation de l'efficacité d'un programme de formation	353
	<i>Diane Berthelette et Luc Desnoyers</i>	
20	Une évaluation des effets d'un programme de réduction des conflits chez les enfants	373
	<i>Françoise Fortin et François Bowen</i>	

21	Une évaluation d'implantation au Burkina Faso avec une méthode mixte	393
	<i>Valéry Ridde, Seni Kouanda, Maurice Yameogo, Kadidiatou Kadio et Aristide Bado</i>	
22	Une évaluation d'implantation et d'efficacité d'un programme destiné aux enfants et à leur famille	415
	<i>Christian Dagenais, Frédéric Nault-Brière, Didier Dupont et Julie Dutil</i>	
	Glossaire	433
	<i>Françoise Jabot et Murielle Bauchet</i>	
	Sigles et abréviations	441
	Bibliographie	443
	Les auteurs	469

**NOUVELLE ÉDITION REVUE ET AUGMENTÉE**

Tous les chapitres de cette nouvelle édition ont été écrits par des pédagogues, des enseignants universitaires et des formateurs rompus depuis de longues années à l'exercice du partage de connaissances en évaluation de programmes, tout en mettant l'accent sur la pratique plutôt que sur la théorie. Nous avons ajouté quatre nouveaux chapitres, car les connaissances en évaluation évoluent constamment, sur la stratégie de l'étude de cas, l'évaluation économique, les approches participatives ou encore l'approche dite réaliste. Il manquait dans la première édition des exemples relatifs à l'usage des méthodes mixtes, décrites dans la première partie. Deux nouveaux chapitres viennent donc combler cette lacune.

Un défi essentiel auquel fait face tout enseignant en évaluation est lié à la maîtrise de la grande diversité des approches évaluatives et des types d'évaluation. La seconde partie de l'ouvrage présente quelques études de cas choisies pour montrer clairement comment les concepts exposés sont employés dans la pratique. Ces chapitres recouvrent plusieurs domaines disciplinaires et proposent divers exemples de pratiques évaluatives.

Valéry Ridde, professeur en santé mondiale, et **Christian Dagenais**, professeur en psychologie, tous deux à l'Université de Montréal, enseignent et pratiquent l'évaluation de programmes au Québec, en Haïti et en Afrique.

AVEC LES TEXTES DE

Aristide Bado
Michael Bamberger
Murielle Bauchet
Diane Berthelette
Pierre Blaise
François Bowen
François Chagnon
Nadia Cunden
Christian Dagenais
Pierre-Marc Daigneault
Luc Desnoyers
Didier Dupont
Julie Dutil
Françoise Fortin

Pierre Fournier
Marie Gervais
Anne Guichard
Robert R. Haccoun
Janie Houle
Françoise Jabot
Steve Jacob
Kadidiatou Kadio
Seni Kouanda
Francine LaBossière
Miri Levin-Rozalis
Isabelle Marcoux
Pierre McDuff
Frédéric Nault-Brière

Bernard Perret
Pierre Pluye
Nancy L. Porteous
Michael Quinn Patton
Valéry Ridde
Émilie Robert
Patricia Rogers
Christine Rothmayr
Jim Rugh
Caroline Tourigny
Josefien Van Olmen
Sophie Witter
Maurice Yameogo
Robert K. Yin

59,95 \$ • 54 €

Disponible en version numérique

www.pum.umontreal.ca

ISBN 978-2-7606-2782-6



9 782760 162782